

Identifying Private Content for Online Image Sharing

by

Ashwini Tonge

B.E., Nagpur University, India, 2010

M.S., University of North Texas, 2017

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Abstract

Images today are increasingly shared online on social networking sites such as Facebook, Flickr, Foursquare, and Instagram. Image sharing occurs not only within a group of friends but also more and more outside a user’s social circles for purposes of social discovery. Despite that current social networking sites allow users to change their privacy preferences, this is often a cumbersome task for the vast majority of users on the Web, who face difficulties in assigning and managing privacy settings. When these privacy settings are used inappropriately, online image sharing can potentially lead to unwanted disclosures and privacy violations. Thus, automatically predicting images’ privacy to warn users about private or sensitive content before uploading these images on social networking sites has become a necessity in our current interconnected world.

In this dissertation, we first explore learning models to automatically predict appropriate images’ privacy as *private* or *public* using carefully identified image-specific features. We study deep visual semantic features that are derived from various layers of Convolutional Neural Networks (CNNs) as well as textual features such as user tags and deep tags generated from deep CNNs. Particularly, we extract deep (visual and tag) features from four pre-trained CNN architectures for object recognition, i.e., AlexNet, GoogLeNet, VGG-16, and ResNet, and compare their performance for image privacy prediction. Results of our experiments on a Flickr dataset of over thirty thousand images show that the learning models trained on features extracted from ResNet outperform the state-of-the-art models for image privacy prediction. We further investigate the combination of user tags and deep tags derived from CNN architectures using two settings: (1) SVM on the bag-of-tags features; and (2) text-based CNN. We compare these models with the models trained on ResNet visual features obtained for privacy prediction.

Further, we present a privacy-aware approach to automatic image tagging, which aims at improving the quality of user annotations, while also preserving the images' original privacy sharing patterns. Experimental results show that, although the user-input tags comprise noise, our privacy-aware approach is able to predict accurate tags that can improve the performance of a downstream application on image privacy prediction, and outperforms an existing privacy-oblivious approach to image tagging. Crowd-sourcing the predicted tags exhibits the quality of our privacy-aware recommended tags.

Finally, we propose an approach for fusing object, scene context, and image tags modalities derived from convolutional neural networks for accurately predicting the privacy of images shared online. Specifically, our approach identifies the set of most competent modalities on the fly, according to each new target image whose privacy has to be predicted. Experimental results show that our approach predicts the sensitive (or private) content more accurately than the models trained on individual modalities (object, scene, and tags) and prior privacy prediction works. Additionally, our approach outperforms the state-of-the-art baselines that also yield combinations of modalities.

Identifying Private Content for Online Image Sharing

by

Ashwini Tonge

B.E., Nagpur University, India, 2010

M.S., University of North Texas, 2017

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Approved by:

Co-Major Professor
Dr. Cornelia Caragea

Approved by:

Co-Major Professor
Dr. Doina Caragea

Copyright

© Ashwini Tonge 2019.

Abstract

Images today are increasingly shared online on social networking sites such as Facebook, Flickr, Foursquare, and Instagram. Image sharing occurs not only within a group of friends but also more and more outside a user’s social circles for purposes of social discovery. Despite that current social networking sites allow users to change their privacy preferences, this is often a cumbersome task for the vast majority of users on the Web, who face difficulties in assigning and managing privacy settings. When these privacy settings are used inappropriately, online image sharing can potentially lead to unwanted disclosures and privacy violations. Thus, automatically predicting images’ privacy to warn users about private or sensitive content before uploading these images on social networking sites has become a necessity in our current interconnected world.

In this dissertation, we first explore learning models to automatically predict appropriate images’ privacy as *private* or *public* using carefully identified image-specific features. We study deep visual semantic features that are derived from various layers of Convolutional Neural Networks (CNNs) as well as textual features such as user tags and deep tags generated from deep CNNs. Particularly, we extract deep (visual and tag) features from four pre-trained CNN architectures for object recognition, i.e., AlexNet, GoogLeNet, VGG-16, and ResNet, and compare their performance for image privacy prediction. Results of our experiments on a Flickr dataset of over thirty thousand images show that the learning models trained on features extracted from ResNet outperform the state-of-the-art models for image privacy prediction. We further investigate the combination of user tags and deep tags derived from CNN architectures using two settings: (1) SVM on the bag-of-tags features; and (2) text-based CNN. We compare these models with the models trained on ResNet visual features obtained for privacy prediction.

Further, we present a privacy-aware approach to automatic image tagging, which aims at improving the quality of user annotations, while also preserving the images' original privacy sharing patterns. Experimental results show that, although the user-input tags comprise noise, our privacy-aware approach is able to predict accurate tags that can improve the performance of a downstream application on image privacy prediction, and outperforms an existing privacy-oblivious approach to image tagging. Crowd-sourcing the predicted tags exhibits the quality of our privacy-aware recommended tags.

Finally, we propose an approach for fusing object, scene context, and image tags modalities derived from convolutional neural networks for accurately predicting the privacy of images shared online. Specifically, our approach identifies the set of most competent modalities on the fly, according to each new target image whose privacy has to be predicted. Experimental results show that our approach predicts the sensitive (or private) content more accurately than the models trained on individual modalities (object, scene, and tags) and prior privacy prediction works. Additionally, our approach outperforms the state-of-the-art baselines that also yield combinations of modalities.

Table of Contents

List of Figures	xii
List of Tables	xvi
Acknowledgements	xviii
1 Introduction	1
1.1 Background	1
1.2 The Importance of Research in Image Privacy Prediction	2
1.3 Motivation and Contributions	3
1.3.1 The Use of “Deep” Features for Online Image Sharing.	4
1.3.2 DeepPrivate Features For Image Privacy Prediction.	5
1.3.3 Privacy-Aware Tag Recommendation for Image Sharing	6
1.3.4 Dynamic Deep Multi-modal Fusion for Image Privacy Prediction.	7
1.4 Dissertation Outline	8
1.4.1 Published Work	11
2 On the Use of “Deep” Features for Online Image Sharing	14
2.1 Introduction	15
2.2 Related work	18
2.3 Convolutional Neural Networks	19
2.4 Image Privacy Classification	20
2.5 Dataset and Evaluation Settings	23
2.6 Experiments and Results	24

2.6.1	Results for Deep Visual Features	25
2.6.2	Results for Deep Image Tags	29
2.7	Chapter Summary and Future Directions	30
3	Image Privacy Prediction Using Deep Neural Networks	31
3.1	Introduction	32
3.2	Related work	36
3.3	Problem Statement	40
3.4	Image encodings	41
3.4.1	Features Derived Through Pre-Trained CNNs	41
3.4.2	Fine-tuned CNN	43
3.4.3	Image Tags (Bag-of-Tags model)	45
3.4.4	Tag CNN	47
3.5	Dataset	48
3.6	Experiments, Results and Observations	49
3.6.1	Classification Experiments for Features Derived From Pre-Trained CNNs	49
3.6.2	The Impact of the CNN Architecture on the Privacy Prediction . . .	51
3.6.3	Fine-Tuned Networks vs. Pre-Trained Networks	54
3.6.4	ResNet Features-Based Models vs. Prior Works	56
3.6.5	Best Performing Visual Features vs. Tag Features	60
3.6.6	Fusion of Visual and Tag Features for Image Privacy Prediction . . .	67
3.7	Chapter Summary and Future Directions	68
4	Privacy-Aware Tag Recommendation for Image Sharing	70
4.1	Introduction	71
4.2	Related work	74
4.2.1	Automatic Image Annotation	74
4.2.2	Tag Recommendation using Collaborative Filtering	78

4.2.3	Online Image Privacy	79
4.3	Privacy-Aware Image Tag Recommendation	82
4.4	Dataset	88
4.5	Experiments and Results	90
4.5.1	Evaluation of Privacy-Aware Recommended Tags by Privacy Prediction	91
4.5.2	Solution to the Cold Start Problem	95
4.5.3	The Proposed Approach vs. Prior Privacy Prediction Works	97
4.5.4	The Proposed Approach vs. Prior Image Annotation Works	99
4.5.5	Quality Assessment of Recommended Tags	101
4.6	Chapter Summary and Future Directions	105
5	Dynamic Deep Multi-modal Fusion for Image Privacy Prediction	107
5.1	Introduction	108
5.2	Related Work	111
5.3	Multi-Modality	113
5.4	Proposed approach	115
5.4.1	Identification of Neighborhoods	116
5.4.2	“Competence” Estimation	117
5.4.3	Dynamic Fusion of Multi-Modality	119
5.5	Dataset	122
5.6	Experiments and Results	123
5.6.1	Impact of Parameters k_v and k_p on DMFP	124
5.6.2	Evaluation of the Proposed Approach	125
5.6.3	Proposed Approach vs. Base Classifiers	127
5.6.4	Proposed Approach vs. Baselines	129
5.6.5	Proposed Approach vs. Prior Image Privacy Prediction Works	132
5.7	Chapter Summary and Future Directions	134

6	Summary and Discussion	135
6.1	Dissertation Summary	135
6.2	Summary of Contributions	138
6.3	Future Directions	142
	Bibliography	144

List of Figures

2.1	Examples of private and public images.	16
2.2	Deep features: 1. CNN is used to extract deep visual features and deep image tags for input images. 2. The features from the fully-connected (fc) layers, “Prob” layer and deep tags are used to predict the class of an image as public or private using SVM. 3. Using these features, we train SVM classifiers to predict the privacy class of an image as <i>private</i> or <i>public</i>	21
2.3	Deep feature representations of a given image.	22
3.1	Examples of images manually identified as private (left) and public (right).	34
3.2	Image encoding using pre-trained CNN: (1) We employ a CNN (e.g. VGG-16) pre-trained on the ImageNet object dataset. (2) We derive high-level features from the image’s visual content using fully connected layers (fc ₆ -V, fc ₇ -V, and fc ₈ -V) and probability layer (softmax) of the pre-trained network. The pictorial representation of VGG is adapted from the image given at: https://github.com/durandtibo/deep_archi_latex	42

3.3	Image encoding using fine-tuned CNN: (1) We modify the last fully-connected layer of the pre-trained network (top network) by changing the output units from 1000 (object categories) to 2 (privacy classes). (2) To train the modified network (bottom network) on privacy dataset, we first adopt weights of all the layers of the pre-trained network as initial weights and then iterate through all the layers using privacy data. (3) To make a prediction for an input image (privacy dataset), we use the probability distribution over 2 privacy classes (softmax layer, yellow rectangle) for the input image obtained by applying the softmax function over the last modified fully-connected layer (fc ₈ -P, bottom network) of the fine-tuned network. The pictorial representation of VGG is adapted from the image given at: https://github.com/durandtibo/deep_archi_latex	44
3.4	Image encoding using tag features: We encode the combination of user tags and deep tags using binary vector representation, showing presence and absence of tags from tag vocabulary V . We set 1 if a tag is present in the tag set or 0 otherwise. We refer this model as Bag-of-Tags (BoT) model.	46
3.5	Tag CNN architecture to classify an image as public or private using image tags.	47
3.6	Performance of various classifiers (LR, NB, RF, SVM) using the features derived from all four architectures AlexNet, GoogLeNet, VGG, and ResNet.	50
3.7	Box plot of F1-measure (overall) obtained for the best-performing features derived from each CNN over five splits.	53
3.8	Precision-recall curves for the private class obtained using features extracted from all four architectures AlexNet (fc ₈), GoogLeNet (loss ₃), VGG-16 (fc ₈ -V) and ResNet (fc-R).	53
3.9	Precision-Recall and Threshold curves for the private class obtained using ResNet features (fc-R) and prior works.	60

3.10	Precision-Recall curves for the private class obtained using visual features (f-R) and tag features as user tags (UT), deep tags (DT-R), the combination of user tags and deep tags (UT + DT-R).	62
3.11	Privacy predictions obtained by image content encodings.	63
3.12	High frequency tag clouds with respect to public and private images.	64
3.13	Tag association graph.	65
3.14	Analysis of top frequently occurring tags.	66
4.1	Anecdotal evidence for visually similar images with privacy-aware user tags.	72
4.2	Illustration of the privacy-aware tag recommendation algorithm using an example: 1) A newly uploaded image on the Web that has an incomplete set of user-input tags, i.e., {"Cute"}, is considered as the target image I . 2) We can use images' tags or content features to compute the similarity between the target image I and the images from the collection \mathcal{D} . For this example, we use visual content features to compute the similarity. 3) Top $r = 3$ tags {"Doll," "Toy," "Coolcat"} are recommended using top $k = 5$ similar images, through our privacy-aware tag recommendation approach. Note that the recommended tags "Doll" and "Toy" are appropriate tags for the target image I and can help correctly characterize its privacy class as <i>public</i>	86
4.3	Tag frequency (%) in the PicAlert dataset. The frequencies are normalized by the size of the dataset.	89
4.4	F1-measure obtained for various parameter values, k and r of Alg. 1. p-Weights and p-Freq are privacy-aware scoring mechanism whereas Weights, Freq and Random are privacy-oblivious scoring mechanisms.	92
4.5	Image with recommended tags, $r=10$	102
4.6	Subjective Adjective (Tags)	105

5.1	Anecdotal evidence of private images and their tags. The feature-level fusion is given as the concatenation of all the features (object, scene, tag) and the decision-level fusion is obtained by averaging the predictions.	109
5.2	Illustration of the proposed approach using an example.	121
5.3	F1-measure obtained for various k_v and k_p values.	125
5.4	Predictions for private images.	128

List of Tables

2.1	Deep visual features vs. Baselines	27
2.2	Class specific privacy prediction performance.	28
2.3	Privacy prediction performance using tag features.	29
3.1	Comparison of SVMs trained on features extracted from pre-trained architectures AlexNet, GoogLeNet, VGG-16 and ResNet. The best performance is shown in bold and blue color. The best performance for each network is shown in italics and orange color.	52
3.2	Fine-tuned networks vs. Pre-trained networks. The best performance is shown in bold and blue color. The performance measures that achieve a better performance after fine-tuning a CNN over pre-trained features are shown in italics and orange color.	54
3.3	Highest performing visual features (fc-R) vs. Prior works.	57
3.4	Visual features vs. Tag features.	61
3.5	Top 50 highly informative tags. We use the combination of deep tags and user tags (DT+UT) to calculate the information gain. User tags are shown in bold.	64
3.6	Results for the combination of visual and tag features.	67
4.1	Privacy-aware weighted sum of tag occurrences ($k = 5$) given that the target image is public. Bold words indicate the top $r = 3$ tags. Since the tag “Cute” appears already in the original set of user tags, we add the next important tag from the ranked list, i.e., “Coolcat.” The tags with same weights are selected randomly.	87
4.2	Datasets summary.	88

4.3	Performance for privacy prediction after adding recommended tags. “ <i>vt</i> ” denotes a set of visible tags and “ <i>rt</i> ” denotes a set of recommended tags, e.g., {“cute”, “toy”, “doll”}. “ <i>r</i> ” is the number of tags recommended.	94
4.4	Visual content-based similarity ($k = 10, r = 5$).	96
4.5	Comparison of privacy prediction performance obtained using the proposed approach and prior privacy prediction approaches.	98
4.6	Privacy-aware Tag recommendation vs. Prior Image Annotation Works. . . .	100
4.7	Gold-standard and User evaluation of privacy-aware and privacy-oblivious recommended tags.	103
4.8	User evaluation of recommended tags that are Noun, Verb, Adjective, and Noun & Verb. Privacy-aware tags are denoted as PA and privacy-oblivious are denoted as PO.	104
5.1	Mathematical notations.	116
5.2	Exploratory analysis.	124
5.3	Evaluation of dynamic multi-modal fusion for privacy prediction (DMFP). . .	125
5.4	Dynamic multi-modal fusion for privacy prediction (DMFP) vs. base classifiers of DMFP.	127
5.5	Errors corrected (%).	128
5.6	Dynamic multi-modal fusion for privacy prediction (DMFP) vs. baselines. . .	131
5.7	Dynamic multi-modal fusion for privacy prediction (DMFP) vs. prior image privacy prediction works.	133

Acknowledgments

First of all, a special thanks to my advisor, Dr. Cornelia Caragea, for her visionary support and unwavering guidance throughout my Ph.D. I am grateful to Dr. Caragea for her encouragement and guidance. I express my sincere gratitude towards my Ph.D. committee members, Dr. Doina Caragea, Dr. Daniel Andresen, and Dr. Caterina Scoglio and the outside chair Dr. Robert Hachiya. Their feedback was valuable and helped me improve my dissertation.

I would like to thank the National Science Foundation for the support from the grant #1421970 to Dr. Cornelia Caragea. The views and conclusions contained in this document should not be interpreted as necessarily representing the official policies, either express or implied, of the National Science Foundation. I am also grateful to all the reviewers for providing insightful comments and directions for additional work.

I am thankful to my colleagues from Kansas State University, University of North Texas and Intel Corporation for their support during my graduate studies. Most importantly, I would like to express my heartfelt gratitude to my parents, Kishor Tonge and Manjusha Tonge, my husband Samir Koppikar who stood firmly beside me in my successes and failures, and my brother Anniruddha Tonge. I would also like to add my gratitude for the incredible support from my in-laws Dilip Koppikar, Medha Koppikar and Sagar Koppikar. This would have not been possible without the love, understanding, patience, and support of my family.

This dissertation is dedicated to all social network users whose privacy needs to be protected using the advancement of artificial intelligence.

Chapter 1

Introduction

In this chapter, we discuss the background and motivation of our study in Image Privacy Prediction.

1.1 Background

Technology today offers innovative ways to share photos with people all around the world, making online photo sharing an incredibly popular activity for Internet users. These users document quotidian details through images and also post pictures of their significant milestones and private events, e.g., family photos and cocktail parties¹. Furthermore, smartphones and other mobile devices facilitate the exchange of information in content sharing sites virtually at any time, in any place. Although current social networking sites allow users to change their privacy preferences, this is often a cumbersome task for the vast majority of users on the Web, who face difficulties in assigning and managing privacy settings². Even though users change their privacy settings to comply with their personal privacy preference, they often misjudge the private information in images, which fails to enforce their own privacy preferences³. Thus, new privacy concerns⁴ are on the rise and mostly emerge due to users' lack of understanding that semantically rich images may reveal sensitive information^{3;5-7}. For example, a seemingly harmless photo of a birthday party may unintentionally

reveal sensitive information about a person’s location, personal habits, and friends.

Along these lines, Gross and Acquisti⁸ analyzed more than 4,000 Carnegie Mellon University students’ Facebook profiles and outlined potential threats to privacy. The authors found that users often provide personal information generously on social networking sites, but they rarely change default privacy settings, which could jeopardize their privacy. Employers often perform background checks for their future employees using social networking sites and about 8% of companies have already fired employees due to their inappropriate social media content⁹. A study carried out by the Pew Research center reported that 11% of the users of social networking sites regret the content they posted¹⁰. The Director of the AI Research at Facebook, Yann LeCun yannlecun urges the development of a digital assistant to warn people about private or sensitive content before embarrassing photos are shared with everyone on social networks.

1.2 The Importance of Research in Image Privacy Prediction

Identifying private or sensitive content from images is inherently difficult because images’ privacy is dependent on the owners’ personality traits and their level of awareness towards privacy. Still, images’ privacy is not purely subjective, but generic patterns of privacy exist. For example, Zerr et al.^{7, 11} conducted a study that manually annotated and consistently rated online images as *private* and *public* by multiple annotators. An image is considered to be private if it belongs to the private sphere (e.g., portraits, family, friends, home) or contains information that can not be shared with everybody on the Web (e.g., private documents), whereas the remaining images are considered to be public⁷.

Researchers showed that generic patterns of images’ privacy can be automatically identified when a large set of images are considered for analysis and investigated binary prediction models based on user tags and image content features such as SIFT (Scale Invariant Feature Transform) and RGB (Red Green Blue)^{7;12;13}. More recently, several studies¹⁴ started

to explore privacy frameworks that leverage the benefits of Convolutional Neural Networks (CNNs) for object recognition since, intuitively, the objects present in images significantly impact images' privacy.

1.3 Motivation and Contributions

Our research is motivated by the fact that increasingly, online users' privacy is routinely compromised by using social and content sharing applications¹⁵. Our models can help users to better manage their participation in online image sharing sites by identifying the sensitive content from the images so that it becomes easier for regular users to control the amount of personal information that they share through these images.

The main purpose of this dissertation is to accurately identify private or sensitive content from images before they are shared on social networking sites. Precisely, given an image, we aim to learn models to classify the image into one of the two classes: *private* or *public*, based on generic patterns of privacy. To achieve our goal, we extract a variety of features from several CNNs and identify those CNNs that have the highest discriminative power for image privacy prediction.

As the privacy of an image can be determined by the presence of one or more objects and scenes described by the visual content and the description associated with it in the form of tags, we consider both visual features and image tags for our analysis. For the purpose of this study, we did not consider other contextual information about images (e.g., personal information about the image owner or the owner social network activities, which may or may not be available or easily accessible) since our goal is to predict the privacy of an image solely from the image's content itself. We rely on the assumption that, although privacy is a subjective matter, generic patterns of images' privacy exist that can be extracted from the images' visual content and textual tags.

In this dissertation, we divided this research into four tasks and they are:

1. The Use of "Deep" Features for Online Image Sharing.

2. DeepPrivate Features For Image Privacy Prediction.
3. Privacy-Aware Tag Recommendation for Image Sharing.
4. Dynamically Fusing Deep Multi-modal Features for Image Sharing.

1.3.1 The Use of “Deep” Features for Online Image Sharing.

We present an analysis of various “deep” feature representations for image privacy prediction (i.e., for predicting the class of an image as *private* or *public*). Unlike previous works, we explore features that can be directly obtained from a pre-trained CNNs for privacy prediction.

Our contributions are as follows:

- We use three deep feature representations corresponding to the output of three fully-connected layers of an eight-layer deep neural network pre-trained on ILSVRC-2012, a subset of ImageNet dataset consisting of 1.2M+ images labeled with 1,000 object categories¹⁶, as well as the probability distribution over the 1,000 categories obtained from the last layer of the network via softmax.
- As discussed earlier, the set of user tags may be incomplete and noisy. Hence, unlike previous works, we leverage CNNs for automatically generating object tags. We also propose the extraction of scene tags to capture additional information from the visual content that is not captured by existing object tags. We call these object and scene tags as “deep tags.”
- We evaluate the performance of the “deep” features (extracted from AlexNet¹⁷) on a subset of the PicAlert dataset of Flickr images, labeled as private or public. The PicAlert dataset was made publicly available by Zerr et al.⁷.
- We empirically show that learning models trained on deep visual features and deep tags for privacy prediction outperform strong baselines such as those trained on hierarchical deep features, SIFT, GIST (global image descriptors) and user provided tags. We also

show that deep visual features provide improved performance for the private class (i.e., correctly identifying more images as private) as compared to baseline approaches.

- Our results show that the deep image tags yield better performing models as compared to user tags and the combination of deep tags and user tags outperforms each set of tags individually.

1.3.2 DeepPrivate Features For Image Privacy Prediction.

Previous studies used only the AlexNet architecture of CNNs for image privacy prediction. To date, many deep CNN architectures have been developed and achieve state-of-the-art performance on object recognition. These CNNs include GoogLeNet¹⁸, VGG-16¹⁹, and ResNet²⁰ (in addition to AlexNet¹⁷). In this task, we present an extensive study to carefully identify the CNN architectures and features derived from these CNNs that can adequately predict the class of an image as *private* or *public*.

Our contributions are as follows:

- We study deep visual semantic features and deep image tags derived from CNN architectures pre-trained on the ImageNet dataset and use them in conjunction with Support Vector Machine (SVM) classifiers for image privacy prediction. Specifically, we extract deep features from four successful (pre-trained) CNN architectures for object recognition, AlexNet, GoogLeNet, VGG-16, and ResNet and compare their performance on the task of privacy prediction. Through carefully designed experiments, we find that ResNet produces the best feature representations for privacy prediction compared with the other CNNs.
- We fine-tune the pre-trained CNN architectures on our privacy dataset and use the softmax function to predict the images' privacy as *public* or *private*. We compare the fine-tuned CNNs with the SVM models obtained on the features derived from the pre-trained CNNs and show that, although the overall performance obtained by the fine-tuned CNNs is comparable to that of SVM models, the fine-tuned networks

provide improved recall for the private class as compared to the SVM models trained on the pre-trained features.

- We show that the best feature representation produced by ResNet outperforms several baselines for image privacy prediction that consider CNN-based models and SVM models trained on traditional visual features such as SIFT and global GIST descriptor.
- Next, we investigate the combination of user tags and deep tags derived from CNNs in two settings: (1) using SVM on the bag-of-tags features; and (2) applying the text CNN²¹ on the combination of user tags and deep tags for privacy prediction using the softmax function. We compare these models with the models trained on the most promising visual features extracted from ResNet (obtained from our study) for privacy prediction. Our results show that the models trained on the visual features perform better than those trained on the tag features.
- Finally, we explore the combination of deep visual features with image tags and show further improvement in performance over the individual sets of features.

1.3.3 Privacy-Aware Tag Recommendation for Image Sharing

We present a privacy-aware approach to automatic image tagging, that aims at improving the quality of user annotations (or user tags), while also preserving the images' original privacy sharing patterns. Precisely, we recommend potential tags for each target image by mining privacy-aware tags from the most similar images of the target image, which we obtain from a large collection of images.

Our contributions are as follows:

- We study our privacy-aware recommended tags obtained by the proposed privacy-aware weighting scheme in an ablation experiment for privacy prediction. In this experiment, we compare various privacy-aware and privacy-oblivious weighting schemes and observe how the privacy prediction performance varies for these weighting schemes. We also experiment with various parameter values to estimate the best parameter setting.

- We compare the performance of privacy prediction using tags recommended by the proposed approach against the tags recommended by a prior state-of-the-art image annotation method. Our objective in this experiment is to verify whether the recommended tags by the proposed approach can capture better privacy characteristics than the prior state-of-the-art annotation.
- We investigate tag recommendation in a binary image privacy prediction task and show that the predicted tags can exhibit relevant cues for specific privacy settings (*public* or *private*) that can be used to improve the image privacy prediction performance.
- Our results show that we achieve a better privacy prediction performance when we add the recommended privacy-aware tags to the original user tags and predicted deep tags of images as compared to prior approaches of image privacy prediction.
- We also evaluate the recommended tags by employing crowd-sourcing to identify relevancy of the suggested tags to images. The results show that, although the user-input tags comprise noise or even some images do not have any tags at all, our approach is able to recommend accurate tags. In addition, we evaluate both privacy-aware and privacy-oblivious recommended tags and show that the privacy-aware recommended tags describe an image’s content more accurately as compared to the privacy-oblivious tags.

1.3.4 Dynamic Deep Multi-modal Fusion for Image Privacy Prediction.

We propose a novel approach that dynamically fuses multi-modal information of online images, derived through Convolutional Neural Networks (CNNs), to adequately identify the sensitive image content. To our knowledge, this is the first study to fuse the most relevant semantic models based on a query image for privacy prediction. In summary, we make the following contributions:

- Our significant contribution is to estimate the competence of object, scene and tag modalities for privacy prediction and dynamically identify the most competent modalities for a target image whose privacy has to be predicted.
- We derive “competence” features from the neighborhood regions of a target image and learn classifiers on them to identify whether a modality is competent to accurately predict the privacy of the target image. To derive these features, we consider privacy and visual neighborhoods of the target image to bring both sensitive and visually similar image content closer.
- We provide an in-depth analysis of our algorithm in an ablation setting, where we record the performance of the proposed approach by removing its various components. The analysis outline the crucial components of our approach.
- Our results show that we identify images’ sensitive content more accurately than single modality models (object, scene, and tag), multi-modality baselines and prior approaches of privacy prediction, depicting that the approach optimally combines the multi-modality for privacy prediction.

1.4 Dissertation Outline

In what follows, we provide a brief description of the chapters in the dissertation. The dissertation is published in AAAI Doctoral Consortium²². Each chapter corresponds to a paper. The research work of the dissertation has been published either in conference proceedings, or is under review in a journal proceeding. Our goal is to accurately identify private or sensitive content from images before they are shared on social networking sites. Precisely, given an image, we aim to learn models to classify the image into one of the two classes: *private* or *public*, based on generic patterns of privacy. This research is motivated by the fact that, increasingly, online users’ privacy is routinely compromised by using social and content sharing applications¹⁵. Identifying sensitive content is inherently difficult because it requires the tool to have an in-depth understanding of the visual content of the image.

Moreover, the problem is very subjective, and users are generally reluctant to give full access to their private images (but only access to the images’ tags) for the image content analysis, which can hinder the personalized privacy prediction using visual features. Hence, we aim to carefully identify features derived from the multi-modal information of the image that can adequately understand the image content and predict the prevalent privacy and sharing needs of users’ uploaded images. The models trained on these features can enable users to better manage their participation in online image sharing systems by making it easier for regular users to control the amount of personal information shared through images, and thus reduce the escalating privacy risks. Moreover, the proposed tags can also provide the relevant cues for privacy-aware image retrieval⁷ and can become an essential tool for surfacing hidden content of the deep Web without exposing sensitive details. In this dissertation, we propose to derive image tags, and visual content features by leveraging CNN architectures, which are used in conjunction with machine learning classifiers and dynamically fuse these modalities to identify sensitive content accurately.

This dissertation is structured as follows:

Chapter 2: We first propose to use the AlexNet CNN architecture¹⁷ to extract deep visual features and deep image tags for all images in the PicAlert dataset⁷ that are labeled as *private* or *public*. AlexNet implements an eight-layer network that is pre-trained on a subset of the ImageNet dataset¹⁶. The first five layers of AlexNet interleave convolution and pooling, whereas the remaining three layers are fully-connected (FC). The convolution layers represent high-level features, whereas the FC layers give the non-linear combination of the features in the layers below. We extract deep visual features from the last three FC layers, and the “prob” layer that produces a probability distribution over 1000 object categories for the input image.

Since, not all images on social networking sites have tags or the set of tags is very sparse²³, we automatically derive tags (deep tags) for images based on their visual content. For deep tags, the top K categories are predicted from the probability distribution extracted from the CNN.²⁴ We propose that scene tags can also contribute along with object tags to learn privacy characteristics of a given image as they can help provide clues into what the

image posters intended to show through the photo. Therefore, we employ two types of deep tags for privacy prediction based on: (1) objects stream, pre-trained on a large scale object dataset (ImageNet)¹⁶, to capture the object information depicted in the image; and (2) scene stream, pre-trained on a large scale scene dataset (Places2)²⁵, to obtain the pattern about scene context of the image²⁶.

Chapter 3: Previous studies used only the AlexNet architecture of CNNs for image privacy prediction. To date, many deep CNN architectures have been developed and achieve state-of-the-art performance on object recognition. These CNNs include GoogLeNet¹⁸, VGG-16¹⁹, and ResNet²⁰ (in addition to AlexNet¹⁷). In this chapter, we present an extensive study to carefully identify the CNN architectures and features derived from these CNNs that can adequately predict the class of an image as *private* or *public*.

Chapter 4: As image tags are at the sole discretion of the users, they tend to be noisy and incomplete. In this chapter, we ask the following questions: *Can we develop an automated approach to recommend accurate image tags that can also take into account the sharing needs of the users for images in questions? Can this method make precise tag recommendations for newly uploaded images that have an incomplete set of user tags or no tags at all? Can these recommended tags help improve the image privacy prediction performance?* We address these questions with our research agenda and propose privacy-aware tag recommendation algorithm, that aims at improving the quality of user annotations while also preserving the images' original sharing settings. These improved set of tags help improve the privacy prediction performance.

Chapter 5: Finally, we propose to combine all the information i.e., object, scene and tags for image privacy prediction and conjecture that simply combining objects, scenes and user tags modalities using feature-level fusion (i.e., concatenation of all object, scene and user tag features) or decision-level fusion (i.e., aggregation of decisions from classifiers trained on objects, scenes and tags) may not always help to identify the sensitive content of images. Thus, we propose a novel approach that dynamically fuses multi-modal information of on-line images (i.e., object, scene and tags), derived through Convolutional Neural Networks (CNNs), to adequately identify the sensitive image content.

Chapter 6: We summarize and conclude the dissertation. We also provide a summary of contributions and directions for future research.

1.4.1 Published Work

- Chapter 2 on *On the Use of “Deep” Features for Online Image Sharing* has been published in Proceedings of the American Association for Artificial Intelligence (AAAI) 2016, 2018^{24;26}, and the Web Conference Companion (WWW) 2018²⁷. To our knowledge, this is the first work to uncover the scene context from the image content for privacy prediction. In this work, we empirically show that learning models trained on deep visual features and deep tags for privacy prediction outperform strong baselines such as those trained on hierarchical deep features, SIFT, GIST (global image descriptors) and user provided tags. We also show that deep visual features provide improved performance for the private class (i.e., correctly identifying more images as private) as compared to baseline approaches. Using “deep” tags, we show that we can achieve performance comparable to the visual content features for privacy prediction. We also evaluate the combination of all three types of tags (object, scene, and user) and show that the combination yields better performance compared with user tags alone and the combination of user tags with scene or object tags.
- Chapter 3 has been submitted to ACM Transactions on the Web (TWEB) and is under review. To our knowledge, this is the first study to provide a detailed analysis of various CNN architectures for privacy prediction. Our comprehensive set of experiments can provide the community with evidence about the best CNN architecture and features for the image privacy prediction task, especially since the results obtained outperformed other complex approaches, on a large dataset of more than 30,000 images. In this work, we extract deep (visual and tag) features from four pre-trained CNN architectures for object recognition, i.e., AlexNet, GoogLeNet, VGG-16, and ResNet, and compare their performance for image privacy prediction. Among all four networks, we observe that ResNet produces the best feature representations for this task. We also

fine-tune the pre-trained CNN architectures on our privacy dataset and compare their performance with the models trained on pre-trained features. The results show that even though the overall performance obtained using the fine-tuned networks is comparable to that of pre-trained networks, the fine-tuned networks provide an improved performance for the private class as compared to models trained on the pre-trained features. Results of our experiments on a Flickr dataset of over thirty thousand images show that the learning models trained on features extracted from ResNet outperform the state-of-the-art models for image privacy prediction. We further investigate the combination of user tags and deep tags derived from CNN architectures using two settings: (1) SVM on the bag-of-tags features; and (2) text-based CNN. We compare these models with the models trained on ResNet visual features obtained for privacy prediction. Our results show that even though the models trained on the visual features perform better than those trained on the tag features, the combination of deep visual features with image tags shows improvements in performance over the individual feature sets.

- Chapter 4 has been published in Proceedings of the ACM conference on Hypertext and Social Media (HT) 2018²⁸. A journal version of this work that augments our study by providing extensive experiments to validate the proposed approach has been accepted to ACM Transactions on Intelligent Systems and Technology (TIST) 2019. In this work, we study our privacy-aware recommended tags obtained by the proposed privacy-aware weighting scheme in an ablation experiment for privacy prediction. In this experiment, we compare various privacy-aware and privacy-oblivious weighting schemes and observe how the privacy prediction performance varies for these weighting schemes. We also experiment with various parameter values to estimate the best parameter setting. We compare the performance of privacy prediction using tags recommended by the proposed approach against the tags recommended by a prior state-of-the-art image annotation method. Our objective in this experiment is to verify whether the recommended tags by the proposed approach can capture better privacy characteristics

than the prior state-of-the-art annotation. We further investigate tag recommendation in a binary image privacy prediction task and show that the predicted tags can exhibit relevant cues for specific privacy settings (*public* or *private*) that can be used to improve the image privacy prediction performance. Our results show that we achieve a better privacy prediction performance when we add the recommended privacy-aware tags to the original user tags and predicted deep tags of images as compared to prior approaches of image privacy prediction. We also evaluate the recommended tags by employing crowd-sourcing to identify relevancy of the suggested tags to images. The results show that, although the user-input tags comprise noise or even some images do not have any tags at all, our approach is able to recommend accurate tags. Additionally, we evaluate both privacy-aware and privacy-oblivious recommended tags and show that the privacy-aware recommended tags describe an image’s content more accurately as compared to the privacy-oblivious tags.

- Chapter 5 has been published in AAAI 2019²⁹ and WWW 2019³⁰. This work identifies the set of most competent modalities on the fly, according to each new target image whose privacy has to be predicted. The approach considers three stages to predict the privacy of a target image, wherein we first identify the neighborhood images that are visually similar and/or have similar sensitive content as the target image. Then, we estimate the competence of the modalities based on the neighborhood images. Finally, we fuse the decisions of the most competent modalities and predict the privacy label for the target image. Experimental results show that our approach predicts the sensitive (or private) content more accurately than the models trained on individual modalities (object, scene, and tags) and prior privacy prediction works. Additionally, our approach outperforms the state-of-the-art baselines that also yield combinations of modalities.

Chapter 2

On the Use of “Deep” Features for Online Image Sharing

Online image sharing in social networking sites such as Facebook, Flickr, and Instagram can lead to unwanted disclosure and privacy violations, when privacy settings are used inappropriately. Despite that social networking sites allow users to set their privacy preferences, this can be cumbersome for the vast majority of users. In this chapter, we explore privacy prediction models for social media that can automatically identify private (or sensitive) content from images, before they are shared online, in order to help protect users’ privacy in social media. More precisely, we study “deep” visual features that are extracted from various layers of a pre-trained deep Convolutional Neural Network (CNN) as well as “deep” image tags generated from the CNN. Experimental results on a Flickr dataset of thousands of images show that the deep visual features and deep image tags can successfully identify images’ private content and substantially outperform previous models for this task.

With the exponential increase in the number of images that are shared online every day, the development of effective and efficient learning methods for image privacy prediction has become crucial. Prior works have used as features automatically derived object tags from images’ content and manually annotated user tags. However, we believe that in addition to objects, the scene context obtained from images’ content can improve the performance

of privacy prediction. Hence, we propose to uncover scene-based tags from images' content using convolutional neural networks. Experimental results on a Flickr dataset show that the scene tags and object tags complement each other and yield the best performance when used in combination with user tags.

2.1 Introduction

The rapid increase in multi-media sharing through social networking sites such as Facebook, Flickr, and Instagram can cause potential threats to users' privacy, when privacy settings are used inappropriately⁵. Many users quickly share private images about themselves, their family and friends, but they rarely change the default privacy settings, which could jeopardize their privacy⁷. These shared images can potentially reveal a user's personal and social habits. Furthermore, the smartphones facilitate the exchange of information virtually at any time with people all around the world. Employers often perform background checks for their future employees using social networking sites and it was reported that about 8% of companies already fired employees due to their indecent media content⁹. A study by the Pew Reserch center¹ reports that 11% of the social networking sites users regret the posted content.

Users' privacy is recognized as a concern by social networking sites researchers as well. For example, the Director of AI Research at Facebook, Yann LeCun² urges the development of a digital assistant, to warn people about sensitive content while uploading embarrassing photos, in order to help them avoid regrets later. Thus, in order to avoid privacy violations and protect users' shared content in social media, it has become critical to develop automated privacy-aware models that can accurately detect private (or sensitive) content from images before they are shared online.

A rule-based classifier that classifies an image as private if it contains people does not work well in a real-world scenario. Consider, for example, an image of a music band in a concert,

¹<http://www.pewinternet.org/2012/02/24/privacy-management-on-social-media-sites>

²<https://www.wired.com/2014/12/fb/all/1>



Figure 2.1: *Examples of private and public images.*

which is identified as consisting of public content. The rule-based classifier will mistakenly classify this image as private. Similarly, an image that does not contain people could be private. Laxton et al.³¹ described a “tele-duplication attack” that allows an adversary to create a physical key duplicate simply from an image. The rule-based model will fail to predict the image of a key as consisting of private (or sensitive) content, which needs to be protected. Figure 2.1 shows examples of *private* and *public* images, i.e., having *private* or *public* content, from a publicly available dataset⁷.

Several studies explored classification models of image privacy using user tags and image content features such as SIFT (or Scale Invariant Feature Transform) or RGB (or Red Green Blue). For example, Zerr et al.⁷ and Squicciarini et al.¹² found that SIFT features and user tags are informative for the task of classifying images as *private* or *public*. Yet, as images’ tags are at the sole discretion of users, they tend to be noisy and incomplete, with many images on the Web containing only a very sparse set of manually annotated tags or no tags at all²³. More recently, due to the success of object recognition from images using Convolutional Neural Networks (CNNs)¹⁷, researchers started to investigate privacy frameworks based on CNNs¹⁴. However, identifying private content automatically is inherently difficult because it requires an in-depth “understanding” of the visual content of the image. In addition, the task is very subjective, depending on factors such as users’ personalities and their privacy awareness. Moreover, users are often reluctant to give access to their private images, which

can hinder the personalized privacy preferences prediction. Recently, Zhong et al.³² discussed challenges faced by both generic and personalized models for image privacy classification. Specifically, they highlight that generic privacy patterns do not capture well an individual’s sharing behavior, whereas personalized models generally require large amounts of individual user data to learn reliable models, and are time and space consuming to train and store models for each user. We recognize that progress should be made on both directions to improve hybrid approaches of generic and personalized models. Thus, in this paper, we aim at identifying a set of generic privacy patterns, i.e., “deep” features (derived from deep CNNs) that have the highest discriminative power for image privacy prediction.

Contributions. We present an analysis of various “deep” feature representations for image privacy prediction (i.e., for predicting the class of an image as *private* or *public*). Unlike previous works, we explore features that can be directly obtained from a pre-trained object CNN for privacy prediction. Specifically, we use three deep feature representations corresponding to the output of three fully-connected layers of an eight-layer deep neural network pre-trained on ILSVRC-2012, a subset of ImageNet dataset consisting of 1.2M+ images labeled with 1,000 object categories¹⁶, as well as the probability distribution over the 1,000 categories obtained from the last layer of the network via softmax. As discussed earlier, the set of user tags may be incomplete and noisy. Hence, unlike previous works, we leverage CNNs for automatically generating object tags. We also propose the extraction of scene tags to capture additional information from the visual content that is not captured by existing object tags. Precisely, we investigate object tags and scene tags, which we call “deep tags” that correspond to the top-ranked probabilities from the probability distribution over the 1,000 object categories and 365 scene categories. We explore the combination of user tags with the object, scene and object-scene tags for privacy prediction. Using these “deep” tags, we show that we can achieve performance comparable to the visual content features for privacy prediction. We evaluate the combination of all three types of tags (object, scene, and user) and show that the combination yields better performance compared with user tags alone and the combination of user tags with scene or object tags. These tags can also provide the relevant cues for privacy-aware image retrieval⁷ and can become an essential tool

for surfacing the hidden content of the deep Web without exposing sensitive details.

We evaluate the performance of the “deep” features (extracted from AlexNet¹⁷) on a subset of the PicAlert dataset of Flickr images, labeled as private or public. The PicAlert dataset was made publicly available by Zerr et al.⁷. We empirically show that learning models trained on deep visual features and deep tags for privacy prediction outperform strong baselines such as those trained on hierarchical deep features, SIFT, GIST (global image descriptors) and user provided tags. We also show that deep visual features provide improved performance for the private class (i.e., correctly identifying more images as private) as compared to baseline approaches. Moreover, the results show that the deep image tags yield better performing models as compared to user tags and the combination of deep tags and user tags outperforms each set of tags individually.

2.2 Related work

Emerging privacy violations and security threats in social media have started to attract various researchers to this field. Several works are carried out to study users’ privacy concerns in social networks, privacy decisions about sharing resources, and the risk associated with them. For example, Ahern et al.⁵ examined privacy decisions and considerations in mobile and online photo sharing. The authors explored critical aspects of privacy such as users’ consideration for privacy decisions, content and context based patterns of privacy decisions, and user behavior towards personal information disclosure. The conclusion was that applications that support and influence the process of users’ privacy decision-making should be developed.

Buschek et al.³³ presented an approach to assigning privacy to shared images using metadata (location, time, shot details) and visual features (faces, colors, edges). Zerr et al.⁷ proposed privacy-aware image classification, and learned classifiers on Flickr photos. Authors considered user-annotated tags and visual features such as color histograms, faces, edge-direction coherence, and SIFT for the privacy classification task and found that SIFT has a high discriminative power for image privacy detection. Consistent with Zerr et al.⁷,

Squicciarini et al. ^{12,13} also found that SIFT and user-annotated tags work best for predicting privacy of users’ images. SIFT as well as GIST are among the most widely used traditional features for image analysis in computer vision. SIFT ³⁴ detects scale, rotation and translation invariant key-points of objects in images and extracts a pool of visual features, which are represented as a “bag-of-visual-words.” GIST ³⁵ encodes global descriptors for images and extracts a set of perceptual dimensions (naturalness, openness, roughness, expansion and ruggedness) that represent the dominant spatial structure of the scene.

Recently, the computer vision community has shifted towards CNNs for tasks such as object detection ^{36;37} and semantic segmentation ³⁸. CNNs have also acquired state of the art results on ImageNet (a highly challenging dataset used for object recognition) ¹⁶ using supervised learning ¹⁷. Karayev et al. ³⁹ described an approach for predicting the style of images using CNNs. Given the recent success of CNNs, Tran et al. ¹⁴ investigated CNNs for privacy prediction and showed improved performance compared with visual features such as SIFT and GIST (this approach is one of our strong baselines). Spyromitros-Xioufis et al. ⁴⁰ explored features extracted from CNNs to provide more accurate personalized privacy classification. Yu et al. ⁴¹ adopted CNNs to achieve semantic image segmentation and also learned object-privacy relatedness to identify privacy-sensitive objects.

2.3 Convolutional Neural Networks

Convolutional Neural Network (CNN) is a type of feed-forward artificial neural network which is inspired by the organization of the animal visual cortex. Learning units in the network are called as neurons. These neurons learn to convert input data i.e. a picture of dog into its corresponding label i.e. “dog” through automated image recognition. Bottom layers of CNN consist of interleaved convolution and pooling layers, and top layers consist of fully-connected (fc) layers, and a probability (prob) layer obtained by applying the softmax function to the input from the previous fc layer, which represents the probability distribution over the available categories for an input image. As we ascend through an architecture, the network acquires: (1) lower layers features (color blobs, lines, corners); (2) middle layer

features (textures resulted from a combination of lower layers); and (3) higher (deeper) layers features (high-level image content like objects obtained by combining middle layers). As online images may contain multiple objects, we consider features extracted from deeper layers as they help to encode the objects precisely.

A CNN exploits the 2D topology of image data, in particular, *local connectivity* through convolution layers, performs *weight sharing* to handle very high-dimensional input data, and can deal with more *abstract or global information* through pooling layers. Each unit within a convolution layer receives as input a small region of its input at location l , denoted $\mathbf{r}_l(\mathbf{x})$ (a.k.a. *receptive field*), and applies a non-linear function to it. More precisely, given an input image \mathbf{x} , a unit that is responsible for region l computes $\sigma(\mathbf{W} \cdot \mathbf{r}_l(\mathbf{x}) + \mathbf{b})$, where \mathbf{W} and \mathbf{b} represent the matrix of weights and the vector of biases, respectively, and σ is a non-linear function such as the sigmoid activation or rectified linear activation function. \mathbf{W} and \mathbf{b} are learned during training and are shared by all units in a convolution layer. Each unit within a pooling layer receives a small region from the previous convolution layer and performs average or max-pooling to obtain more abstract features. During training, layers in CNNs are responsible for a forward pass and backward pass. The forward pass takes inputs and generates the outputs. The backward pass takes gradients with respect to the output and computes the gradient with respect to the parameters and to the inputs, which are consecutively back-propagated to the previous layers⁴².

2.4 Image Privacy Classification

The privacy of an image can be determined by the presence of one or more objects described by the visual content and the description associated with it in the form of tags.

Problem Statement: Given an image to be uploaded online, the task is to classify it into one of the two classes: *private* or *public*, i.e., consisting of private or public content, respectively.

Next, we describe the features used in the classification.

Feature Extraction: We extract “deep” features from images using a pre-trained CNN.

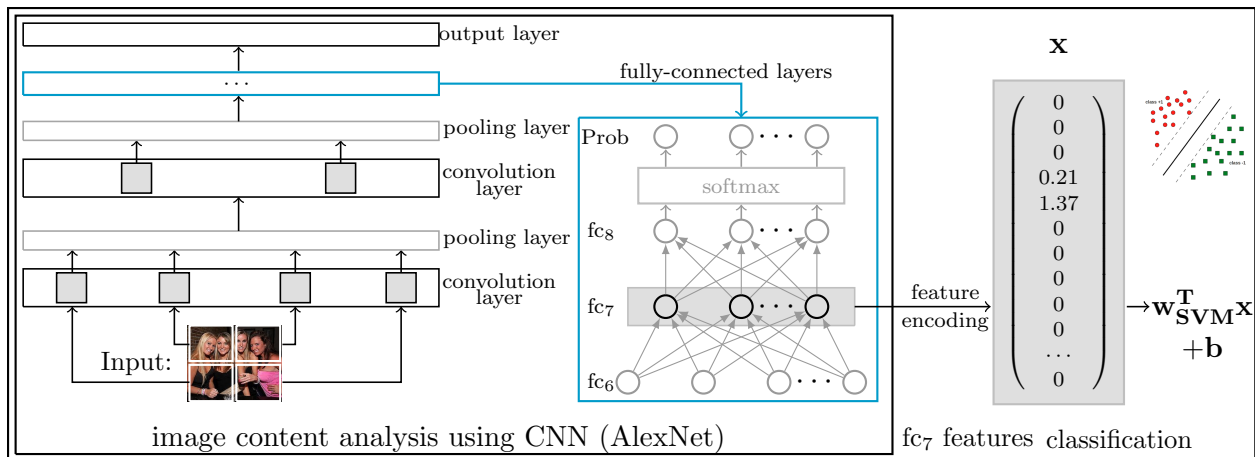


Figure 2.2: Deep features: 1. CNN is used to extract deep visual features and deep image tags for input images. 2. The features from the fully-connected (fc) layers, “Prob” layer and deep tags are used to predict the class of an image as public or private using SVM. 3. Using these features, we train SVM classifiers to predict the privacy class of an image as private or public.

A CNN is a feed-forward neural network in which bottom layers consist of interleaved convolution and pooling layers, and top layers consist of fully-connected (FC) layers, a probability (prob) layer obtained by applying the softmax function to the input from the previous FC layer, and finally the output layer, which outputs the probabilities of the objects in the input image. This is illustrated in Figure 2.2. A CNN exploits the 2-dimensional topology of image data, e.g., *local connectivity* through convolution layers, performs *weight sharing* to handle very high-dimensional input data, and can deal with more *abstract* or *global information* through pooling layers. The convolution layers represent high-level features of images, whereas the FC layers give the non-linear combination of the features in the layers below. In CNNs, features are extracted from images through each layer in a feed-forward fashion. We used the AlexNet CNN architecture¹⁷ to extract deep visual features and deep image tags for all images in the dataset, which are labeled as *private* or *public*. AlexNet implements an eight-layer network pre-trained on the ILSVRC-2012 object classification subset of the ImageNet dataset¹⁶. The first five layers of AlexNet interleave convolution and pooling layers, and the last three layers are fully-connected. We show the AlexNet architecture in Figure 2.2. The reason for using features derived from a pre-trained network is that the sensitive

content is limited for model training and training or fine-tuning a deep network requires a large amount of privacy data.

Deep Visual Features: We extracted deep visual features from the last three fully-connected layers, which are referred as fc_6 , fc_7 , and fc_8 , and from the “prob” layer (the cyan block in Figure 2.2). Figure 2.3 shows the visual features fc_6 , fc_7 , fc_8 and Prob extracted using AlexNet. The dimensions of fc_6 , fc_7 , and fc_8 are 4096, 4096 and 1000, respectively, and the “prob” layer produces a probability distribution over 1000 object categories for the input image. The conditional probability distribution over object categories c can be defined using a softmax function as given below:

$$P(y = c|\mathbf{z}) = \frac{\exp(z_k)}{\sum_j \exp(z_j)}$$

where, in our case, \mathbf{z} is the output of the last fully connected layer (i.e., the fc_8 layer).

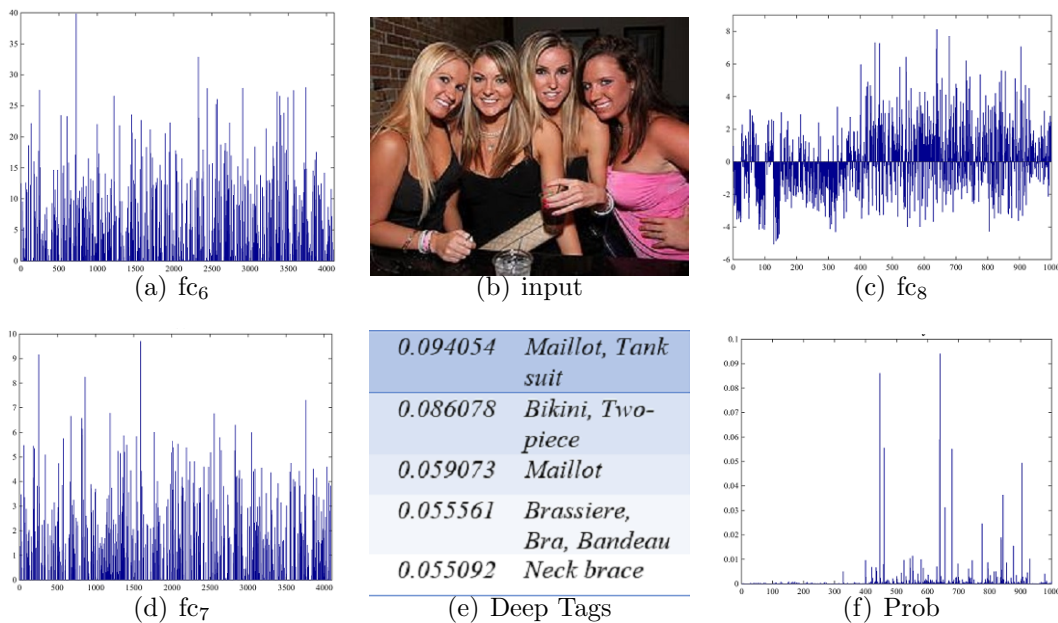


Figure 2.3: Deep feature representations of a given image.

Deep Image Tags: It is interesting to mention that not all images on social networking sites have tags or the set of tags is very sparse²³. Thus, we use an automatic annotation technique to derive tags for images based on their visual content. We believe that scene

tags can contribute along with object tags to learn privacy characteristics of a given image, as they can help provide clues into what the image owners intended to show through the photo. Therefore, we employ two types of semantic features for privacy prediction based on: (1) objects stream, pre-trained on a large scale object dataset (ImageNet)¹⁶, to capture the objects depicted in the image; and (2) scene stream, pre-trained on a large scale scene dataset (Places2)²⁵, to obtain the pattern about scene context of the image. For automatic image annotation, we predict the top K object/scene categories from the probability distribution extracted from the deep neural network. More precisely, given an input image \mathbf{x} , the automatically annotated tags (or deep tags) correspond to the object/scene categories of the top K probabilities.

Object-centric Tags. To automatically obtain object tags from the visual content, we adopt an approach given in²⁴. We use the probability distribution over 1000 object categories for the input image obtained by applying the softmax function over the last fully-connected layer of the AlexNet Convolutional Neural Network (CNN)¹⁷. AlexNet is pre-trained on the ImageNet dataset and obtained from the CAFFE distribution⁴². We consider the top k objects of highest probabilities as *object tags*. We can see from Figure 2.3 that the deep tags such as “Maillot” and “Tank suit” are representative for the image (note that only top $K = 5$ deep tags are shown), but important tags such as “people” and “women” are not included. This is because the 1,000 object categories used for training do not contain these tags.

Scene-centric Tags. Similar to object tags, we obtain the top k scenes derived from the probability distribution over 365 scene categories of the pre-trained AlexNet on the Places2 dataset. We refer to the top k predicted scenes as *scene tags*.

2.5 Dataset and Evaluation Settings

We trained and evaluated models based on deep features on a subset of 32,000 Flickr images sampled from the PicAlert dataset, made available by Zerr et al.⁷. PicAlert consists of Flickr images on various subjects, which are manually labeled as *private* or *public* by external

viewers. The data have been labeled by six teams providing a total of 81 users of ages between 10 and 59 years of varied backgrounds. In our experiments, the 32,000 images are split into **Train** and **Test** sets of 10,000 and 22,000 images, respectively. We consider a higher number of test images (compared to Training images) to evaluate the “deep” features on a large set of unseen images for limited number of training images. Each experiment was repeated five times with a different train/test split (obtained using five different random seeds). The results presented in the next section are averaged across the five runs. Also, the F1-score is calculated as a weighted average of the F1-score of both the classes. The public and private images are in the ratio of 3:1 in both train and test.

Evaluation Setting. To evaluate the deep features, we used the Support Vector Machine (SVM) classifier implemented in Weka and chose the hyper-parameters that gave the best performance on the **Train** set using 10-fold cross-validation. We experimented with $C = \{0.001, 0.01, 1.0, \dots, 10.0\}$, kernels: Polynomial and RBF, the γ parameter in RBF, and the degree d of a polynomial. Hyper-parameters shown in all result tables follow the format: “R/P,C, γ / d ” where “R” denotes “RBF” and “P” denotes “Polynomial.”

2.6 Experiments and Results

In this section, we present the experimental evaluation of the deep features. We compare the performance of the models trained on deep visual features with the models trained on baseline visual features for privacy prediction. In previous works^{7:12}, tag features perform very well for privacy prediction, and hence, we examine the quality of tag features using both user annotated tags and automatically annotated (deep) tags. In order to learn the tags which help in privacy prediction, we also perform an analysis of the most informative tags with respect to the binary privacy settings.

2.6.1 Results for Deep Visual Features

Experimental design: We wish to identify the most promising visual features from the set of deep features that have the highest discriminative ability for privacy classes. To achieve this, we first compare the deep visual features among each other. We then compare the performance of models based on deep visual features with several baselines that we described below.

Baselines. Tran et al.¹⁴ proposed PCNH, a privacy CNN-based framework, that combines features obtained from two architectures: one that extracts convolutional features (size = 24), and another that extracts object features (size = 24). The Object CNN is a deep network of 11 layers obtained by appending three fully-connected layers of size 512, 512, 24 at the end of the fully-connected layer of AlexNet. The PCNH framework is first trained on the ImageNet dataset and then fine-tuned on a small privacy dataset. As images' privacy greatly depends on the objects in images, we believe that the features controlling the distinct attributes of the objects obtained through the higher number of neurons (4096 neurons in FC₇ of AlexNet) can better approximate the privacy function compared with adding more non-linear layers (as in PCNH). The increase in the number of complex non-linear layers introduces more parameters to learn, and at the same time, with comparatively small amount of training data (PicAlert vs. ImageNet), which may result in over-fitting. Moreover, training such a deep network on ImageNet and then fine-tuning on the privacy data significantly increases the processing power and time complexity. Furthermore, if new objects are added to the object dataset, the networks need to be retrained from scratch. Conversely, features derived from state-of-the-art CNN architectures can reduce the overhead of re-training and still achieve good performance for privacy prediction. Hence, we compare models trained on the deep features extracted from the fully-connected layers of AlexNet, as explained in Section 3, with the PCNH privacy framework, and consider the latter as our first baseline. Unlike Tran et al.¹⁴ who used 800 images in their evaluation, we evaluate our models on a large set of images (22000) to validate the performance of the deep features for a large variety of image subjects.

We regard classifiers trained on the best performing features between SIFT, GIST, and their combination as the second strong baseline. Our choice of these features is motivated by their good performance over other features such as colors, patterns, and edge directions in prior works^{7;12}. For SIFT, we constructed a vocabulary of 128 visual words for our experiments as visual words of length 500 or above did not yield significant improvement over 128 visual words. For a given image, GIST is computed by: (1) convolving the image with 32 Gabor filters at 4 scale and 8 orientations, which produces 32 feature maps; (2) dividing the feature map into a 4×4 grid and averaging feature values of each cell; (3) concatenating these 16 averaged values for 32 feature maps, which result in a feature vector of 512 (16×32) length.

We also compare the performance of the deep features with two naive rule-based classifiers, which predict an image as *private* if it contains persons. Otherwise, the image is classified as *public*. For the first rule-based classifier, we detect front and profile faces by using Viola-Jones algorithm⁴³. For the second rule-based classifier, we consider user tags such as “women,” “men,” “people.” Recall that these tags are not present in the set of 1,000 categories of ILSVRC-2012, and hence, we restrict to user tags only. If an image contains one of these tags, we consider it as “private,” otherwise “public.” For the deep visual features, we use the AlexNet pre-trained CNN implemented in CAFFE⁴², which is an open-source framework for deep neural networks. We resize images in both **Train** and **Test** to the CAFFE convolutional neural net compatible size of 227×227 and encode each image using the three deep feature representations corresponding to the output of the layers FC_6 , FC_7 , FC_8 , and “Prob,” which is the probability distribution obtained from FC_8 via softmax.

Results: Table 2.1 shows results of the comparison (Precision, Recall, F1- Measure and Accuracy) of SVMs using each deep feature type extracted from AlexNet, fc_6 , fc_7 , fc_8 , and “Prob,” and the results of their comparison with the performance of baselines (i.e., SVMs trained using the baseline features), on **Test**. We can see from the table that the SVMs trained on fc_7 and fc_8 perform similarly, and the performance improves as we go from fc_6 to fc_7 . This is because higher layers of the network capture high level feature descriptions of objects present in the image. We notice that all fc_6 , fc_7 , fc_8 deep features are able to achieve

Features	H-Param	Acc %	F1	Prec	Re
#1 Deep visual features					
fc ₆	R,1.0,0.05	85.49	0.844	0.847	0.855
fc ₇	R,2.0,0.01	85.83	0.851	0.851	0.858
fc ₈	R,1.0,0.05	85.80	0.851	0.851	0.858
Prob	R,5.0,1.0	83.18	0.824	0.822	0.832
#2 Hierarchical Deep Features ¹⁴					
PCNH	R,1.0,0.01	84.21	0.833	0.832	0.842
#3 SIFT/GIST (Zerr et al. ⁷ , Squicciarini et al. ^{12, 13})					
SIFT	P,1.0,2.0	77.31	0.674	0.598	0.773
GIST	R,0.001,0.5	77.33	0.674	0.598	0.773
SIFT+GIST	R,0.05,0.5	72.67	0.704	0.691	0.727
#4 Rule-based models					
Rule-1	–	77.35	0.683	0.694	0.672
Rule-2	–	77.93	0.673	0.704	0.644

Table 2.1: *Deep visual features vs. Baselines*

performance higher than 85% in terms of all compared measures. Note that a naive baseline which classifies every image as “public” obtains an accuracy of 75%. It is worth mentioning that “prob” features perform worse than the features extracted from the fully-connected layers. One possible explanation could be that squashing the values at the previous layer (e.g., fc₈ in AlexNet) through the softmax function, which yields the “prob” layer, produces a non-linearity that is less useful for SVM compared to the un-transformed values. The results of fully-connected layers over the “prob” layer are statistically significant for p-values < 0.05.

Table 2.1 shows also that deep visual features fc₆, fc₇, fc₈ provide better feature representations than baseline visual features for privacy prediction. Precisely, the models obtained using deep visual features extracted from AlexNet outperform models trained on baseline features, PCNH, SIFT, GIST and SIFT + GIST. For example, F1-measure improves from 0.833 obtained by PCNH features to 0.851 obtained by fc₈. We achieve improvement in F1-measure as high as 15% over SIFT + GIST models, i.e., our second baselines. “Prob” features also perform better than SIFT + GIST. With a paired T-test, our improvements over the baseline approaches for F1-measure are statistically significant for p-values < 0.05.

It is also interesting to note that rules based on facial features exhibit better performance

	Private			Public		
Features	F1	Prec	Re	F1	Prec	Re
#1 Deep visual features						
fc7	0.642	0.752	0.56	0.912	0.88	0.946
#2 Hierarchical Deep Features (Tran et al. ¹⁴)						
PCNH	0.598	0.708	0.518	0.902	0.869	0.937
#3 SIFT/GIST (Zerr et al. ⁷ , Squicciarini et al. ^{12, 13})						
SIFT+GIST	0.27	0.343	0.223	0.832	0.793	0.874
#4 Rule-based models						
Rule-1	0.509	0.47	0.556	0.853	0.875	0.832
Rule-2	0.458	0.373	0.593	0.897	0.914	0.88

Table 2.2: *Class specific privacy prediction performance.*

than SIFT and GIST and suggest that features representing persons are helpful to predict private content of images. This is consistent with Tran et al. ¹⁴, who showed that adding the “person” category in the object classes helped to improve the accuracy. However, AlexNet-based “deep” features outperform: (1) the rule-based models based on facial features by more than 10% in terms of all measures (see Table 2.1, #4 Rule-based models), and (2) the PCNH features that incorporate the “person” category by more than 2.5-3% in terms of all measures (see Table 2.1, #3 Hierarchical Deep Features). Simple rule-based models will not suffice for this task and advanced AI technology for image content analysis such as deep learning is required.

We also show the class specific privacy prediction performance in Table 2.2 to identify which features characterize the private class effectively as sharing private images on the Web with everyone is not desirable. We found that the SVMs trained on AlexNet-based deep visual features obtain improved performance for the private class as compared with the SVMs trained on the baseline features. Precisely, using the best-performing deep visual features FC₇, F1-measure for the private class improves from 0.598 obtained by PCNH to 0.642 obtained by FC₇. Similarly, the F1-measure for the public class improves from 0.902 obtained by PCNH to 0.912 for FC₇.

Next, we examine the quality of tag features and contrast the deep image tags with the user annotated tags.

Feat.	$k = 2$					$k = 10$				
	Acc %	F1	Pre.	Re.	#Inc	Acc %	F1	Pre.	Re.	#Inc
UT	81.73	0.789	0.803	0.817	-	81.73	0.789	0.803	0.817	-
UT + ST	82.26	0.797	0.81	0.823	293	83.21	0.814	0.821	0.832	503
UT + OT	83.09	0.812	0.819	0.831	477	84.35	0.833	0.834	0.843	755
UT + ST + OT	83.59	0.819	0.825	0.836	587	84.80	0.841	0.84	0.848	854

Table 2.3: Privacy prediction performance using tag features.

2.6.2 Results for Deep Image Tags

Experimental design: We investigate the performance of SVMs on user tags and deep image tags for privacy prediction. We also examine the combination of user tags and deep tags, which captures different aspects of an image.

For user tags, we remove special characters and numbers from the user tags, as they do not provide any information with respect to privacy. Examples of user tags for the image in Figure 2.3 are: “Birthday Party,” “Night Life,” “People,” etc. To obtain object and scene tags (deep tags) from CNNs, we experimented with two values of k as $k = 2$ and $k = 10$ (for the top k tags). The choice for $k = 2$ is motivated by the fact that an image may contain only a few scenes or objects, whereas the choice for $k = 10$ is motivated by its best results. We also contrast the combination of user, scene, and object tags with the combination of user and scene tags and user tags alone. To encode the automatically derived scene and object tags, we use the probability of the tag obtained from the softmax layer of the corresponding CNN. The user tags are encoded using a binary representation.

Results: Table 2.3 shows the results obtained from the experiments for tag features on the **Test** and compares the performance obtained using models trained on deep tags, user tags and their combination. Precisely, the table shows the performance obtained before and after adding scene tags (ST), object tags (OT) and scene + object tags (ST+OT) to the user tags (UT). We observe that models trained on the combination of all tag types yield the best performance and show an improvement as high as 5.2% in F1-measure over models trained on UT alone. From the table, we also notice that deep tags (object and scene tags)

perform better than user tags, however, the combination of the two outperforms each one individually, the user tags and the deep tags (UT + OT and UT + ST). This can be justified by the fact that the user tags have some general tags, whereas deep tags contain some specific tags, which capture various aspects of the data. To see this, using only general tags can cause overlap in the two different privacy classes. For example, if we consider more general tags such as “clothes” instead of “swimsuit,” then the tag can appear in both classes and hence will fail to differentiate between them. Similarly, if we would consider only very specific tags, the models may overfit and will not generalize well on unseen data. Moreover, we note that the combination of UT+ST performs better than UT alone, but does not perform as good as the combination of OT+UT. Table 2.3 also shows the increase in the number of accurate predictions (denoted by #Inc) for UT+ST, UT+OT, and UT+ST+OT over the user tags. As can be seen, the highest increase is achieved by the combination of UT+ST+OT.

2.7 Chapter Summary and Future Directions

In this chapter, we explored AI technology, i.e., deep features extracted from various CNN layers, for image privacy classification. Our results show that the deep visual features corresponding to the fully-connected layers of the AlexNet CNN outperform those corresponding to the “prob” layer. We also examined user annotated tags and deep tags (generated from the “prob” layer) and found that the combination of both the tags outperforms individual sets of tags. In addition, models trained on deep features yield improvement in performance over several baselines. The result of our classification task is expected to aid other very practical applications. For example, a law enforcement agent who needs to review digital evidence on a suspected equipment to detect sensitive content in images and videos, e.g., child pornography. The learning models developed here can be used to filter or narrow down the number of images and videos having sensitive or private content before other more sophisticated approaches can be applied to the data.

In future, other CNN architectures can be explored for privacy prediction. Also, user tags can be explored in various ways, e.g., to include information from description, comment, anchor tags to obtain additional information about the image.

Chapter 3

Image Privacy Prediction Using Deep Neural Networks

Images today are increasingly shared online on social networking sites such as Facebook, Flickr, Foursquare, and Instagram. Image sharing occurs not only within a group of friends but also more and more outside a user’s social circles for purposes of social discovery. Despite that current social networking sites allow users to change their privacy preferences, this is often a cumbersome task for the vast majority of users on the Web, who face difficulties in assigning and managing privacy settings. When these privacy settings are used inappropriately, online image sharing can potentially lead to unwanted disclosures and privacy violations. Thus, automatically predicting images’ privacy to warn users about private or sensitive content before uploading these images on social networking sites has become a necessity in our current interconnected world.

In this chapter, we explore learning models to automatically predict appropriate images’ privacy as *private* or *public* using carefully identified image-specific features. We study deep visual semantic features that are derived from various layers of Convolutional Neural Networks (CNNs) as well as textual features such as user tags and deep tags generated from deep CNNs. Particularly, we extract deep (visual and tag) features from four pre-trained CNN architectures for object recognition, i.e., AlexNet, GoogLeNet, VGG-16, and ResNet,

and compare their performance for image privacy prediction. Among all four networks, we observe that ResNet produces the best feature representations for this task. We also fine-tune the pre-trained CNN architectures on our privacy dataset and compare their performance with the models trained on pre-trained features. The results show that even though the overall performance obtained using the fine-tuned networks is comparable to that of pre-trained networks, the fine-tuned networks provide an improved performance for the private class as compared to models trained on the pre-trained features. Results of our experiments on a Flickr dataset of over thirty thousand images show that the learning models trained on features extracted from ResNet outperform the state-of-the-art models for image privacy prediction. We further investigate the combination of user tags and deep tags derived from CNN architectures using two settings: (1) SVM on the bag-of-tags features; and (2) text-based CNN. We compare these models with the models trained on ResNet visual features obtained for privacy prediction. Our results show that even though the models trained on the visual features perform better than those trained on the tag features, the combination of deep visual features with image tags shows improvements in performance over the individual feature sets. Our code, features, and the dataset used in experiments are available at <https://github.com/ashwinitonge/deepprivate.git>.

3.1 Introduction

Online image sharing through social networking sites such as Facebook, Flickr, and Instagram is on the rise, and so is the sharing of private or sensitive images, which can lead to potential threats to users' privacy when inappropriate privacy settings are used in these platforms. Many users quickly share private images of themselves and their family and friends, without carefully thinking about the consequences of unwanted disclosure and privacy violations^{5:7}. For example, it is common now to take photos at cocktail parties and share them on social networking sites without much hesitation. The smartphones facilitate the sharing of photos virtually at any time with people all around the world. These photos can potentially reveal a user's personal and social habits and may be used in the detriment of the photos' owner.

Gross and Acquisti⁸ analyzed more than 4,000 Carnegie Mellon University students' Facebook profiles and outlined potential threats to privacy. The authors found that users often provide personal information generously on social networking sites, but they rarely change default privacy settings, which could jeopardize their privacy. In a parallel study, Lipford et al.² showed that, although current social networking sites allow users to change their privacy preferences, the vast majority of users on the Web face difficulties in assigning and managing privacy settings. Interestingly, Orekondy et al.³ showed that, even when users change their privacy settings to comply with their personal privacy preference, they often misjudge the private information in images, which fails to enforce their own privacy preferences. Not surprising, employers these days often perform background checks for their future employees using social networks and about 8% of companies have already fired employees due to their inappropriate social media content⁹. A study carried out by the Pew Research center reported that 11% of users of social networks regret the posted content¹⁰. The Director of the AI Research at Facebook, LeCun¹ urges the development of a digital assistant to warn people about private or sensitive content before embarrassing photos are shared with everyone on social networks.

Identifying private or sensitive content from images is inherently difficult because images' privacy is dependent on the owners' personality traits and their level of awareness towards privacy. Still, images' privacy is not purely subjective, but generic patterns of privacy exist. Consider, for example, the images shown in Figure 3.1, which are manually annotated and consistently rated as *private* and *public* by multiple annotators in a study conducted by Zerr et al.^{7,11}. Notice that the presence of people generally pinpoints to private images, although this is not always true. For example, an image of a musical band in concert is considered to be public. Similarly, images with no people in them could be private, e.g., images with door keys, music notes, legal documents, or someone's art are considered to be private. Indeed, Laxton et al.³¹ described a "tele-duplication attack" that allows an adversary to create a physical key duplicate simply from an image.

Researchers showed that generic patterns of images' privacy can be automatically identified when a large set of images are considered for analysis and investigated binary prediction



Figure 3.1: *Examples of images manually identified as private (left) and public (right).*

models based on user tags and image content features such as SIFT (Scale Invariant Feature Transform) and RGB (Red Green Blue)^{7;12;13}. More recently, several studies^{14;24;27} started to explore privacy frameworks that leverage the benefits of Convolutional Neural Networks (CNNs) for object recognition since, intuitively, the objects present in images significantly impact images’ privacy (as can be seen from Figure 3.1). However, these studies used only the AlexNet architecture of CNNs on small dataset sizes. To date, many deep CNN architectures have been developed and achieve state-of-the-art performance on object recognition. These CNNs include GoogLeNet¹⁸, VGG-16¹⁹, and ResNet²⁰ (in addition to AlexNet¹⁷). Towards this end, in this chapter, we present an extensive study to carefully identify the CNN architectures and features derived from these CNNs that can adequately predict the class of an image as *private* or *public*. Our research is motivated by the fact that increasingly, online users’ privacy is routinely compromised by using social and content sharing applications¹⁵. Our models can help users to better manage their participation in online image sharing sites by identifying the sensitive content from the images so that it becomes easier for regular users to control the amount of personal information that they share through these images.

Our contributions are as follows:

- We study deep visual semantic features and deep image tags derived from CNN architectures pre-trained on the ImageNet dataset and use them in conjunction with Support Vector Machine (SVM) classifiers for image privacy prediction. Specifically, we extract deep features from four successful (pre-trained) CNN architectures for object recogni-

tion, AlexNet, GoogLeNet, VGG-16, and ResNet and compare their performance on the task of privacy prediction. Through carefully designed experiments, we find that ResNet produces the best feature representations for privacy prediction compared with the other CNNs.

- We fine-tune the pre-trained CNN architectures on our privacy dataset and use the softmax function to predict the images' privacy as *public* or *private*. We compare the fine-tuned CNNs with the SVM models obtained on the features derived from the pre-trained CNNs and show that, although the overall performance obtained by the fine-tuned CNNs is comparable to that of SVM models, the fine-tuned networks provide improved recall for the private class as compared to the SVM models trained on the pre-trained features.
- We show that the best feature representation produced by ResNet outperforms several baselines for image privacy prediction that consider CNN-based models and SVM models trained on traditional visual features such as SIFT and global GIST descriptor.
- Next, we investigate the combination of user tags and deep tags derived from CNNs in two settings: (1) using SVM on the bag-of-tags features; and (2) applying the text CNN²¹ on the combination of user tags and deep tags for privacy prediction using the softmax function. We compare these models with the models trained on the most promising visual features extracted from ResNet (obtained from our study) for privacy prediction. Our results show that the models trained on the visual features perform better than those trained on the tag features.
- Finally, we explore the combination of deep visual features with image tags and show further improvement in performance over the individual sets of features.

3.2 Related work

Emerging privacy violations in social networks have started to attract various researchers to this field¹⁵. Researchers also provided public awareness of privacy risks associated with images shared online^{44;45}. Along this line, several works are carried out to study users' privacy concerns in social networks, privacy decisions about sharing resources, and the risk associated with them^{8;46-51}.

Moreover, several works on privacy analysis examined privacy decisions and considerations in mobile and online photo sharing^{5;8;52;53}. For example, Ahern et al.⁵ explored critical aspects of privacy such as users' consideration for privacy decisions, content and context based patterns of privacy decisions, and how different users adjust their privacy decisions and behavior towards personal information disclosure. The authors concluded that applications that could support and influence users' privacy decision-making process should be developed. Jones and O'Neill⁵² reinforced the role of privacy-relevant image concepts. For instance, the authors determined that people are more reluctant to share photos capturing social relationships than photos taken for functional purposes; certain settings such as work, bars, concerts cause users to share less. Besmer and Lipford⁵³ mentioned that users want to regain control over their shared content, but meanwhile, they feel that configuring proper privacy settings for each image is a burden.

More recent and related to our line of work are the automated image privacy approaches that have been explored along four lines of research: *social group based approaches*, in which users' profiles are used to partition the friends' lists into multiple groups or circles, and the friends from the same circle are assumed to share similar privacy preferences; *location-based approaches*, in which location contexts are used to control the location-based privacy disclosures; *tag-based approaches*, in which tags are used for privacy setting recommendations; and *visual-based approaches*, in which the visual content of images is leveraged for privacy prediction.

Social group based approaches.

Several works emerged to provide the automated privacy decisions for images shared online based on the social groups or circles^{7;54–69}. For example, Christin et al.⁵⁵ proposed an approach to share content with the users within privacy bubbles. Privacy bubbles represent the private sphere of the users and the access to the content is provided by the bubble creator to people within the bubble. Bonneau et al.⁶⁰ introduced the notion of privacy suites which recommend users a set of privacy settings that “expert” users or the trusted friends have already established so that ordinary users can either directly accept a setting or perform minor modifications only. Fang and LeFevre⁶¹ developed a privacy assistant to help users grant privileges to their friends. The approach takes as input the privacy preferences for the selected friends and then, using these labels, constructs a classifier to assign privacy labels to the rest of the (unlabeled) friends based on their profiles. Danezis⁶³ generated privacy settings based on the policy that the information produced within the social circle should remain in that circle itself. Along these lines, Adu-Oppong et al.⁷⁰ obtained privacy settings by forming clusters of friends by partitioning a user’s friends’ list. Yuan et al.⁶⁹ proposed an approach for context-dependent and privacy-aware photo sharing. This approach uses the semantics of the photo and the requester’s contextual information in order to define whether an access to the photo will be granted or not at a certain context. These social group based approaches mostly considered the user trustworthiness, but ignored the image content sensitiveness, and thus, they may not necessarily provide appropriate privacy settings for online images as the privacy preferences might change according to sensitiveness of the image content.

Location-based approaches.

These approaches^{69;71–81} leverage geo-tags, visual landmarks and other location contexts to control the location-based privacy disclosures. The geo-tags can be provided manually via social tagging or by adding location information automatically through the digital cameras or smart-phones having GPS. The location can also be inferred by identifying places from

the shared images through the computer vision techniques.

Tag-based approaches.

Previous work in the context of tag-based access control policies and privacy prediction for images^{7;54;56;62;67;68;82–88} showed initial success in tying user tags with access control rules. For example, Squicciarini et al.^{62, 88}, Zerr et al.⁷, and Vyas et al.⁸⁴ explored learning models for image privacy prediction using user tags and found that user tags are informative for predicting images’ privacy. Moreover, Squicciarini et al.⁶⁷ proposed an Adaptive Privacy Policy Prediction framework to help users control access for their shared images. The authors investigated social context, image content, and metadata as potential indicators of privacy preferences. Klemperer et al.⁶⁸ studied whether the user annotated tags help to create and maintain access-control policies more intuitively. However, the scarcity of tags for many online images²³ and the dimensions of user tags precluded an accurate analysis of images’ privacy. Hence, in our previous work,^{24;26–28} we explored automatic image tagging and showed that the predicted tags combined with user tags can improve the overall privacy prediction performance.

Visual-based approaches.

Several works used visual features derived from the images’ content and showed that they are informative for predicting images’ privacy settings^{7;12–14;24;27;33;89–102}. For example, Buschek et al.³³ presented an approach to assigning privacy to shared images using metadata (location, time, shot details) and visual features (faces, colors, edges). Zerr et al.⁷ proposed privacy-aware image classification and learned classifiers on Flickr images. The authors considered image tags and visual features such as color histograms, faces, edge-direction coherence, and Scale Invariant Feature Transform (SIFT) for the privacy classification task. SIFT as well as GIST are among the most widely used traditional features for image analysis in computer vision. SIFT³⁴ detects scale, rotation, and translation invariant key-points of objects in images and extracts a pool of visual features, which are represented as a “bag-of-

visual-words.” GIST³⁵ encodes global descriptors for images and extracts a set of perceptual dimensions (naturalness, openness, roughness, expansion, and ruggedness) that represent the dominant spatial structure of the scene. Squicciarini et al.^{12,13} performed an in-depth analysis of image privacy classification using Flickr images and found that SIFT and image tags work best for predicting privacy of users’ images.

Recently, the computer vision community has shifted towards convolutional neural networks (CNNs) for tasks such as object detection^{36;37} and semantic segmentation³⁸. CNNs have acquired state-of-the-art results on ImageNet for object recognition¹⁶ using supervised learning¹⁷. Given the recent success of CNNs, several researchers^{14;24;27;89–91} showed promising privacy prediction results compared with visual features such as SIFT and GIST. Yu et al.⁴¹ adopted CNNs to achieve semantic image segmentation and also learned object-privacy relatedness to identify privacy-sensitive objects.

Using CNNs, some works started to explore personalized privacy prediction models^{3;32;40}. For example, Spyromitros-Xioufis et al.⁴⁰ used features extracted from CNNs to provide personalized image privacy classification. Zhong et al.³² proposed a Group-Based Personalized Model for image privacy classification in online social media sites that learns a set of archetypical privacy models (groups) and associates a given user with one of these groups. Orekondy et al.³ defined a set of privacy attributes, which were first predicted from the image content and then used these attributes in combination with users’ preferences to estimate personalized privacy risk. Although there is evidence that individuals’ sharing behavior is unique, Zhong et al.³² argued that personalized models generally require large amounts of user data to learn reliable models, and are time and space consuming to train and store models for each user, while taking into account possible sudden changes of users’ sharing activities and privacy preferences. Orekondy et al.³ tried to resolve some of these limitations by clustering users’ privacy profiles and training a single classifier that maps the target user into one of these clusters to estimate the personalized privacy score. However, the users’ privacy profiles are obtained using a set of attributes, which are defined based on the Personally Identifiable Information¹⁰³, the US Privacy Act of 1974 and official online social network rules, instead of collecting opinions about sensitive content from the actual users of social networking sites.

Hence, the definition of sensitive content may not meet a user’s actual needs, which limits their applicability in a real-world usage scenario¹⁰⁴. In this context, it is worth mentioning that CNNs were also used in another body of privacy related work such as multi-party privacy conflict detection¹⁰⁵ and automatic redaction of sensitive image content⁹⁷.

The image representations using visual features and tags are pivotal in above privacy prediction works. In this chapter, we aim to study “deep” features derived from CNNs, by abstracting out users’ privacy preferences and sharing behavior. Precisely, our goal is to identify a set of “deep” features that have the highest discriminative power for image privacy prediction and to flag images that contain private or sensitive content before they are shared on social networking sites. To our knowledge, this is the first study to provide a detailed analysis of various CNN architectures for privacy prediction. Our comprehensive set of experiments can provide the community with evidence about the best CNN architecture and features for the image privacy prediction task, especially since the results obtained outperformed other complex approaches, on a large dataset of more than 30,000 images.

3.3 Problem Statement

Our goal is to accurately identify private or sensitive content from images before they are shared on social networking sites. Precisely, given an image, we aim to learn models to classify the image into one of the two classes: *private* or *public*, based on generic patterns of privacy. Private images belong to the private sphere (e.g., self-portraits, family, friends, someone’s home) or contain information that one would not share with everyone else (e.g., private documents). Public images capture content that can be seen by everyone without incurring privacy violations. To achieve our goal, we extract a variety of features from several CNNs and identify those features that have the highest discriminative power for image privacy prediction.

As the privacy of an image can be determined by the presence of one or more objects described by the visual content and the description associated with it in the form of tags, we consider both visual features and image tags for our analysis. For the purpose of this study,

we did not consider other contextual information about images (e.g., personal information about the image owner or the owner social network activities, which may or may not be available or easily accessible) since our goal is to predict the privacy of an image solely from the image’s content itself. We rely on the assumption that, although privacy is a subjective matter, generic patterns of images’ privacy exist that can be extracted from the images’ visual content and textual tags.

We describe the feature representations considered for our analysis in the next section.

3.4 Image encodings

In this section, we provide details on visual content encodings and tag content encodings derived from various CNNs (pre-trained and fine-tuned) to carefully identify the most informative feature representations for image privacy prediction. Particularly, we explore four CNN architectures, AlexNet¹⁷, GoogLeNet¹⁸, VGG-16¹⁹, and ResNet²⁰ to derive features for all images in our dataset, which are labeled as private or public. The choice of these architectures is motivated by their good performance on the large scale ImageNet object recognition challenge¹⁶. We also leverage a text-based CNN architecture used for sentence classification²¹ and apply it to images’ textual tags for privacy prediction.

3.4.1 Features Derived Through Pre-Trained CNNs

We describe a diverse set of features derived from CNN architectures pre-trained on the ILSVRC-2012 object classification subset of the ImageNet dataset that contains 1000 object categories and 1.2 million images¹⁶. We consider powerful features obtained from various fully-connected layers of a CNN that are generated by the previous convolutional layers, and use them to learn a decision function whose sign represents the class (*private* or *public*) assigned to an input image \mathbf{x} . The activations of the fully connected layers capture the complete object contained in the region of interest. Hence, we use the activations of the fully-connected layers of a CNN as a feature vector. For image encoding, we also use the

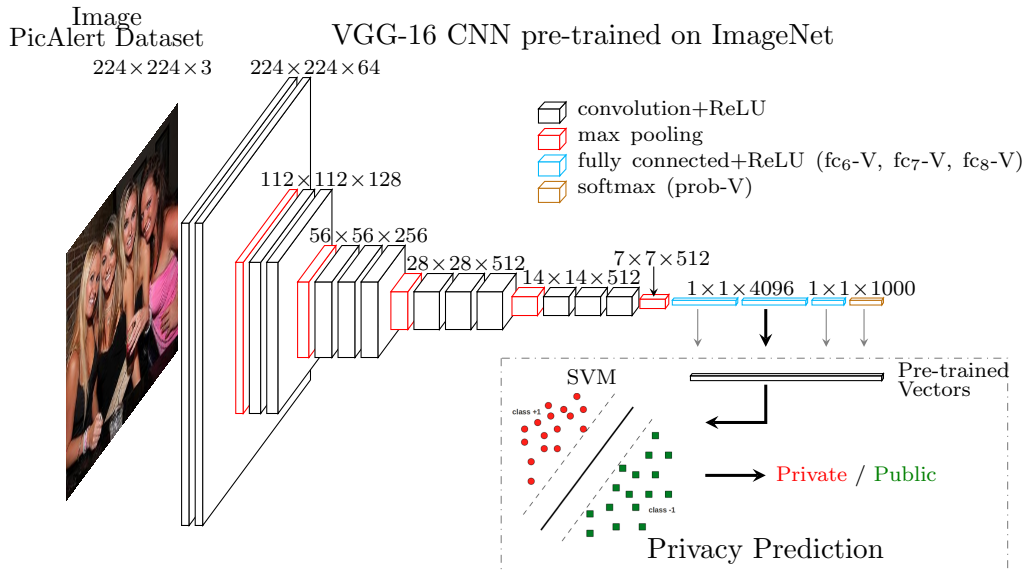


Figure 3.2: Image encoding using pre-trained CNN: (1) We employ a CNN (e.g. VGG-16) pre-trained on the ImageNet object dataset. (2) We derive high-level features from the image’s visual content using fully connected layers (fc_6-V , fc_7-V , and fc_8-V) and probability layer (softmax) of the pre-trained network. The pictorial representation of VGG is adapted from the image given at: https://github.com/durandtibo/deep_archi_latex.

probability (prob) layer obtained by applying the softmax function to the output of the (last) fully-connected layer. We extract features from the four pre-trained CNNs as follows.

The **AlexNet** architecture implements an eight-layer network; the first five layers of AlexNet are convolutional, and the remaining three layers are fully-connected. We extract features from the three fully-connected layers, which are referred as fc_6-A , fc_7-A , and fc_8-A , and from the output layer denoted as “prob-A.” The dimensions of fc_6-A , fc_7-A , fc_8-A , and prob-A are 4096, 4096, 1000, and 1000, respectively.

The **GoogLeNet** architecture implements a 22 layer deep network with Inception architecture. The architecture is a combination of all layers with their output filter bank concatenated so as to form input for the next stage. We extract features from the last two layers named as “loss₃-G/classifier” (InnerProduct layer) and the output layer denoted as “prob-G.” The dimension of loss₃-G and prob-G is 1000.

The **VGG-16** architecture implements a 16 layer deep network; a stack of convolutional layers with a very small receptive field: 3×3 followed by fully-connected layers. The ar-

chitecture contains 13 convolutional layers and 3 fully-connected layers. The number of channels of the convolutional layers starts from 64 in the first layer and then increases by a factor of 2 after each max-pooling layers until it reaches 512. We refer to features extracted from the fully-connected layers as fc_6-V , fc_7-V , fc_8-V , and the output layer as “prob-V.” The dimensions of fc_6-V , fc_7-V , fc_8-V , and prob-V are 4096, 4096, 1000, and 1000, respectively.

The **ResNet** (or Residual network) alleviates the vanishing gradient problem by introducing short paths to carry gradient throughout the extent of very deep networks and allows the construction of deeper architectures. A residual unit with an identity mapping is defined as:

$$X^{l+1} = X^l + \mathcal{F}(X^l)$$

where X^l is the input and X^{l+1} is the output of the residual unit; \mathcal{F} is a residual function, e.g., a stack of two 3×3 convolution layers in²⁰. The main idea of the residual learning is to learn the additive residual function \mathcal{F} with respect to X^l ¹⁰⁶. Intuitively, ResNets can be explained by considering residual functions as paths through which information can propagate easily. This interprets as ResNets learn more complex feature representations which are combined with the shallower descriptions obtained from previous layers. We refer to features extracted from the fully-connected layer as fc-R and the output layer as “prob-R.” The dimension of fc-R and prob-R is 1000.

The feature extraction using the pre-trained network for an input image from our dataset is shown in Figure 3.2. In the figure, we show VGG-16 as the pre-trained network for illustrating the feature extraction.

3.4.2 Fine-tuned CNN

For this type of encoding, models trained on a large dataset (e.g., the ImageNet dataset) are fine-tuned using a smaller dataset (e.g., the privacy-labeled dataset). Fine-tuning a network is a procedure based on the concept of transfer learning^{107;108}. This strategy fine-tunes the weights of the pre-trained network by continuing the back-propagation on the small dataset, i.e., privacy dataset in our scenario. The features become more dataset-specific after fine-

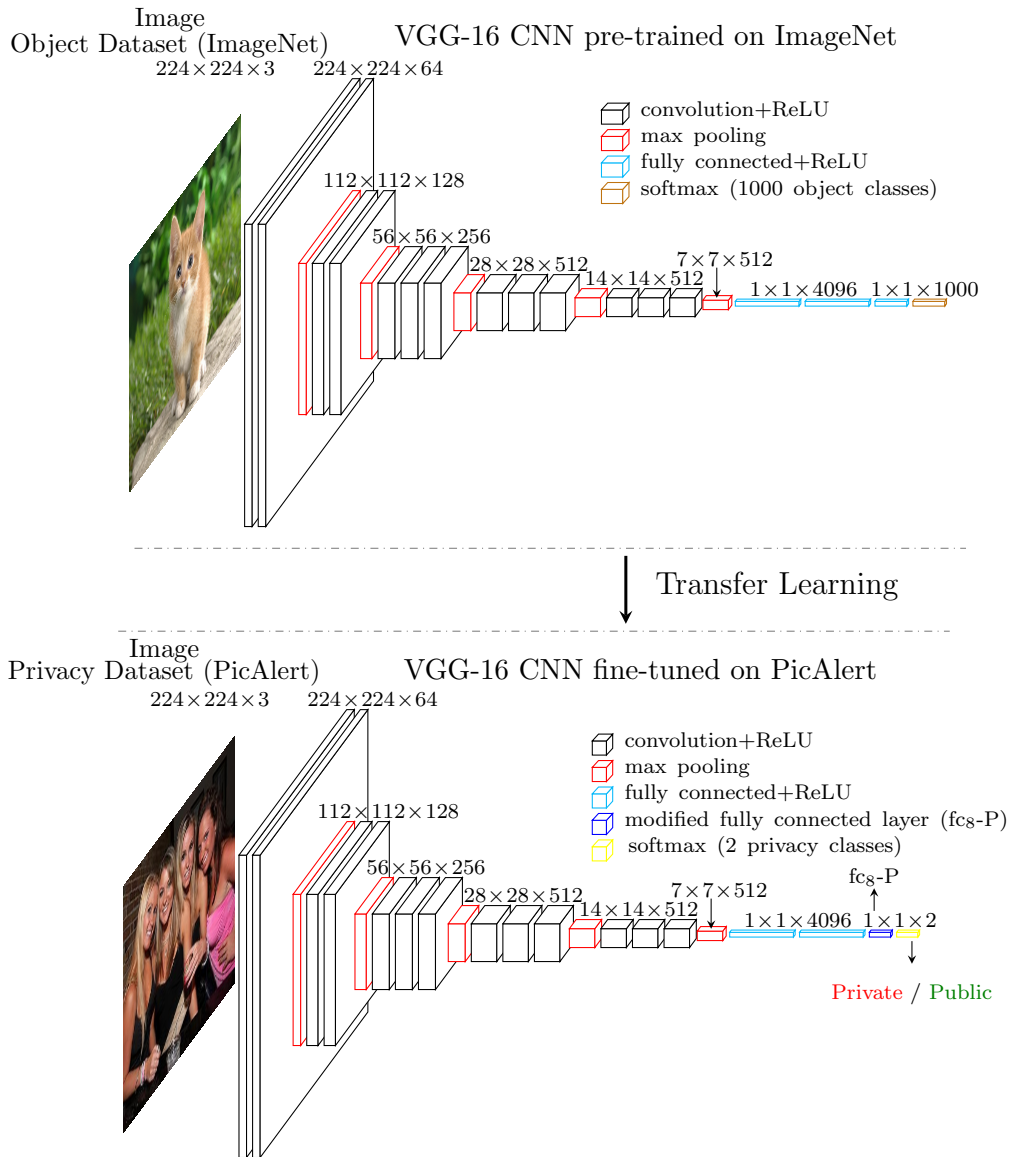


Figure 3.3: Image encoding using fine-tuned CNN: (1) We modify the last fully-connected layer of the pre-trained network (top network) by changing the output units from 1000 (object categories) to 2 (privacy classes). (2) To train the modified network (bottom network) on privacy dataset, we first adopt weights of all the layers of the pre-trained network as initial weights and then iterate through all the layers using privacy data. (3) To make a prediction for an input image (privacy dataset), we use the probability distribution over 2 privacy classes (softmax layer, yellow rectangle) for the input image obtained by applying the softmax function over the last modified fully-connected layer (fc_8-P , bottom network) of the fine-tuned network. The pictorial representation of VGG is adapted from the image given at: https://github.com/durandtibo/deep_archi_latex.

tuning, and hence, are distinct from the features obtained from the pre-trained CNN. We modify the last fully-connected layer of all four network architectures, AlexNet, GoogLeNet, VGG-16, and ResNet by changing the output units from 1000 (object categories) to 2 (with respect to privacy classes) (e.g., changing fc_8 with 1000 output units to fc_8 -P with 2 output units). We initialize the weights of all the layers of this modified architectures with the weights of the respective layers obtained from the pre-trained networks. We train the network by iterating through all the layers of the networks using the privacy data. We use the softmax function to predict the privacy of an image. Precisely, we use the probability distribution over 2 privacy classes for the input image obtained by applying the softmax function over the modified last fully-connected layer (e.g. fc_8 -P in VGG-16) of the fine-tuned networks (See Figure 3.3, second network, blue rectangle). The conditional probability distribution over 2 privacy classes can be defined using a softmax function as given below:

$$P(y = P_r|\mathbf{z}) = \frac{\exp(z_{P_r})}{\exp(z_{P_u}) + \exp(z_{P_r})}, P(y = P_u|\mathbf{z}) = \frac{\exp(z_{P_u})}{\exp(z_{P_u}) + \exp(z_{P_r})}$$

where, in our case, \mathbf{z} is the output of the modified last fully connected layer (e.g., the fc_8 -P layer of VGG-16) and P_r and P_u denote *private* and *public* class, respectively. The fine-tuning process using VGG-16 is shown in Figure 3.3.

3.4.3 Image Tags (Bag-of-Tags model)

Prior works on privacy prediction^{7;12;24;88} found that the tags associated with images are indicative of their sensitive content. Tags are also crucial for image-related applications such as indexing, sharing, searching, content detection and social discovery^{109–112}. Since not all images on social networking sites have user tags or the set of user tags is very sparse²³, we use an automatic technique to annotate images with tags based on their visual content as described in our previous work²⁴. Precisely, we predict top k object categories from the probability distribution extracted from a pre-trained CNN. These top k categories are images’ deep tags, used to describe an image. For example, we obtain deep tags such as

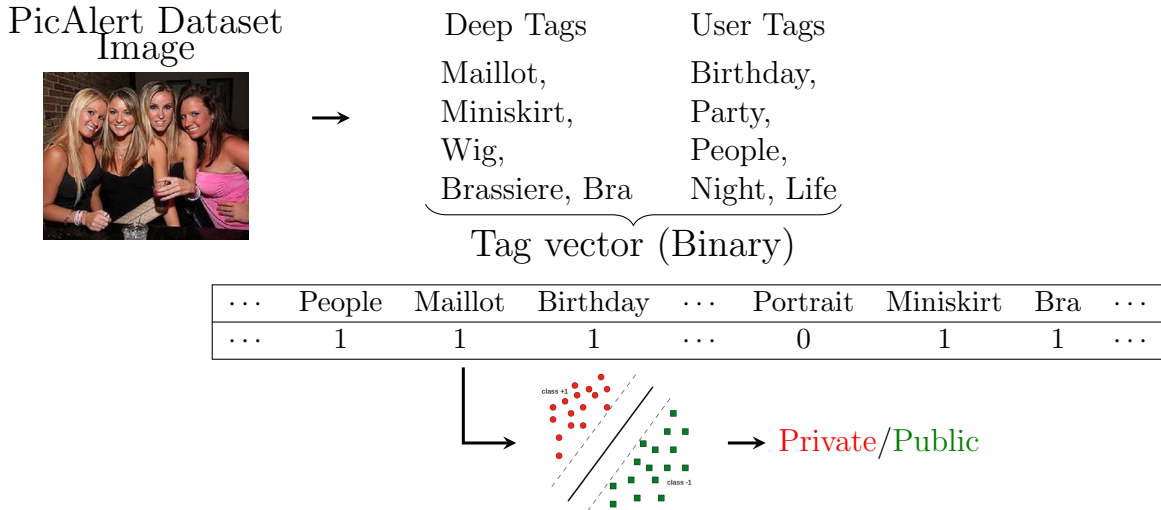


Figure 3.4: *Image encoding using tag features: We encode the combination of user tags and deep tags using binary vector representation, showing presence and absence of tags from tag vocabulary V . We set 1 if a tag is present in the tag set or 0 otherwise. We refer this model as Bag-of-Tags (BoT) model.*

“Maillot,” “Wig,” “Brassiere,” “Bra,” “Miniskirt” for the picture in Figure 3.4 (note that only top 5 deep tags are shown in the figure). Note that the deep tags give some description about the image, but still some relevant tags such as “people” and “women” are not included since the 1000 object categories of the ImageNet dataset do not contain these tags. Images on social networking sites also give additional information about them through the tags assigned by the user. We call these tags “User Tags.” Examples of user tags for the image in Figure 3.4 are: “Birthday Party,” “Night Life,” “People,” etc. For user tags, we remove special characters and numbers from the user tags, as they do not provide any information with respect to privacy.

We combine deep tags and user tags and generate a binary vector representation for the tag set of an image, illustrating presence or absence of tags from tag vocabulary V . Particularly, we create a vector of size $|V|$, wherein, for all tags in the tag set, we set 1 on the position of the tag in the vocabulary (V) and 0 otherwise. We refer to this model as a Bag-of-Tags (BoT) model and show its pictorial representation in Figure 3.4.

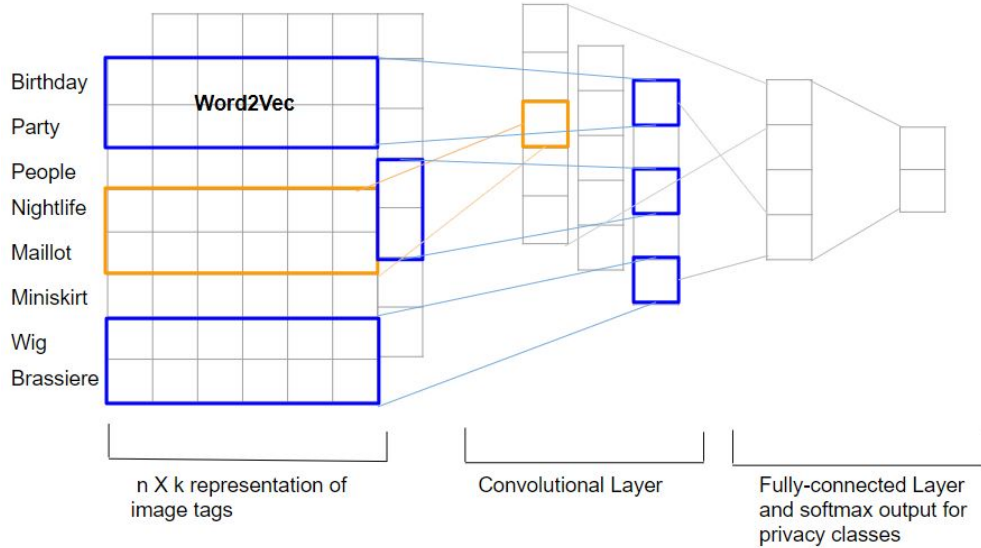


Figure 3.5: *Tag CNN architecture to classify an image as public or private using image tags.*

3.4.4 Tag CNN

CNN based models have achieved exceptional results for various NLP tasks such as semantic parsing¹¹³, search query retrieval, sentence modeling¹¹⁴, sentence classification²¹, and other traditional NLP tasks¹¹⁵. Kim²¹ developed a CNN architecture for sentence level classification task. A sentence contains keywords in the form of objects, subjects, and verbs that help in the classification task. Image tags are nothing but keywords that are used to describe an image. Thus, for privacy prediction, we employ a CNN architecture that has proven adequate for sentence classification²¹.

The CNN architecture by Kim²¹ shown in Figure 3.5 is a slight variant of the CNN architecture of Collobert et al.¹¹⁵. This architecture contains one layer of convolution on top of word vectors obtained from an unsupervised neural language model. The first layer embeds words (tags in our case) into the word vectors. The word vectors are first initialized with the word vectors that were trained on 100 billion words of Google News, given by Le and Mikolov¹¹⁶. Words that are not present in the set of pre-trained words are initialized randomly. These word vectors are then fine-tuned on the tags from the privacy dataset. The next layer performs convolutions on the embedded word vectors using multiple filter sizes of

3, 4 and 5, where we use 128 filters from each size and produce a tag feature representation. A max-pooling operation¹¹⁵ over a feature map is applied to take the maximum value of the features to capture the most important feature of each feature map. These features are passed to a fully connected softmax layer to obtain the probability distribution over privacy labels. An illustration of the Tag CNN model can be seen in Figure 3.5.

3.5 Dataset

We evaluated our approach on a subset of 32,000 Flickr images sampled from the PicAlert dataset, made available by Zerr et al.^{7,11}. PicAlert consists of Flickr images on various subjects, which are manually labeled as *public* or *private* by external viewers. The dataset contains photos uploaded on Flickr during the period from January to April 2010. The data have been labeled by six teams providing a total of 81 users of ages between 10 and 59 years. One of the teams included graduate computer science students working together at a research center, whereas other teams contained users of social platforms. Users were instructed to consider that their camera has taken these pictures and to mark them as “private,” “public,” or “undecidable.” The guideline to select the label is given as private images belong to the private sphere (like self-portraits, family, friends, someone’s home) or contain information that one would not share with everyone else (such as private documents). The remaining images are labeled as public. In case no decision could be made, the image was marked as undecidable. Each image was shown to at least two different users. In the event of disagreement, the photos were presented to additional users. We only consider images that are labeled as public or private.

For all experiments, our 32,000 images dataset is split into train and test sets of 27,000 and 5,000 images, respectively. Each experiment is repeated five times with a different train/test split (obtained using five different random seeds), and the final results are averaged across the five runs. The public and private images are in the ratio of 3:1 in both train and test sets.

3.6 Experiments, Results and Observations

In this section, we perform a broad spectrum of experiments to evaluate features extracted from various deep architectures in order to understand which architecture can capture the complex privacy characteristics and help to distinguish between privacy classes. We first choose the machine learning classifier between generative models, ensemble methods, and discriminative algorithms for privacy prediction. Then, we use the chosen classifier to examine the visual features extracted from all four deep architectures AlexNet, GoogLeNet, VGG-16, and ResNet pre-trained on object data. We further investigate these architectures by fine-tuning them on the privacy data. Next, we compare the performance of models trained on the highest performing features with that of the state-of-the-art models and baseline approaches for privacy prediction. Additionally, we show the performance of the deep tags obtained through all four pre-trained networks and also study the combination of deep tags and user tags in details for privacy prediction. We show the tag performance in two settings: (1) Bag-of-Tags models and (2) Tag CNN. We analyze the most promising features derived from both visual and tag encodings for privacy classification. We also provide a detailed analysis of the most informative tags for privacy prediction. Finally, we show the performance of the models trained on the fusion of visual and most informative tag features.

3.6.1 Classification Experiments for Features Derived From Pre-Trained CNNs

We first determine the classifier that works best with the features derived from the pre-trained CNNs. We study the performance of the features using the following classification algorithms: Naive Bayes (NB), Random Forest (RF), Logistic Regression (LR) and Support Vector Machine (SVM). NB is a generative model, whereas RF is an ensemble method using decision trees, and SVM and LR are discriminative algorithms. We evaluate the performance of these classifiers using the features derived from the last fully-connected layer of all the architectures, i.e., fc_8 -A of AlexNet, $loss_3$ -G of GoogLeNet, fc_8 -V of VGG-16, and fc -R of

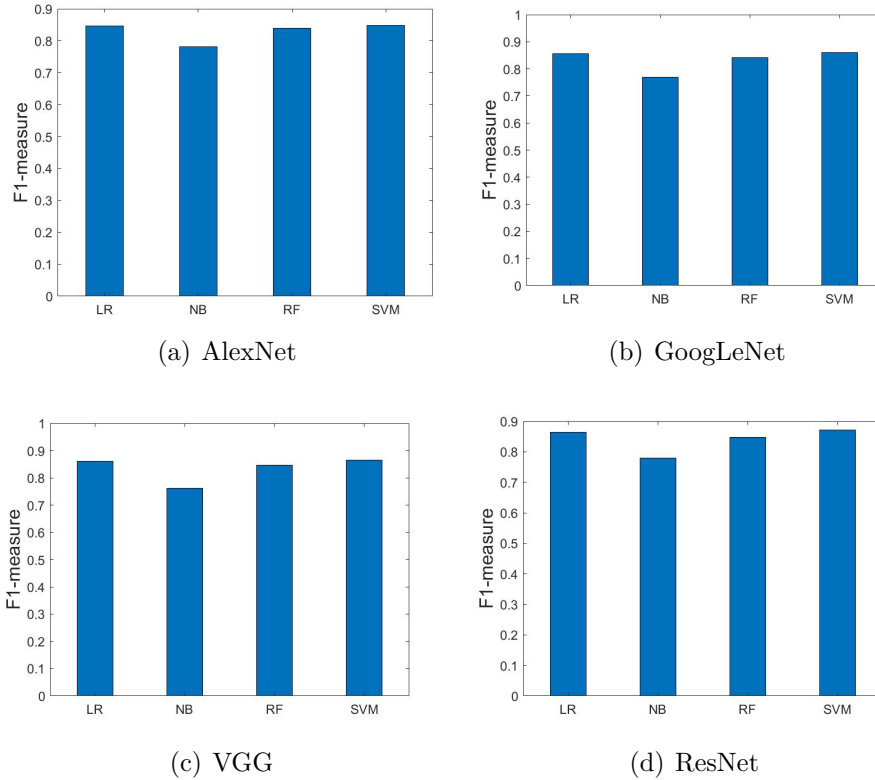


Figure 3.6: Performance of various classifiers (*LR*, *NB*, *RF*, *SVM*) using the features derived from all four architectures *AlexNet*, *GoogLeNet*, *VGG*, and *ResNet*.

ResNet. Figure 3.6 shows the performance of these classifiers in terms of F1-measure for all four architectures. From the figure, we notice that almost all the classifiers perform similarly except NB which performs worse. For example, for Alexnet, with NB we get an F1-measure of 0.781, whereas SVM obtains an F1-measure of 0.849. We can also observe that, generally, SVM and LR perform better than RF. For example, for ResNet, using SVM, we get an F1-measure of 0.872, whereas with RF we get an F1-measure of 0.848. SVM and LR perform comparably to each other for almost all the architectures except for ResNet. For ResNet, we obtain F1-measure of 0.872 and 0.865 using SVM and LR, respectively. The results of SVM over the LR classifier are statistically significant for p-values < 0.05 . Thus, we chose to use SVM with the features derived from pre-trained CNNs for all of our next experiments.

To evaluate the proposed features, we used the SVM Weka implementation and chose the hyper-parameters that gave the best performance using 10-fold cross-validation on the training set. We experimented with $C = \{0.001, 0.01, 1.0, \dots, 10.0\}$, kernels: Polynomial

and RBF, the γ parameter in RBF, and the degree d of a polynomial. Hyper-parameters shown in all subsequent tables follow the format: “R/P,C, γ / d ” where “R” denotes “RBF” and “P” denotes “Polynomial” kernel.

3.6.2 The Impact of the CNN Architecture on the Privacy Prediction

In this experiment, we aim to determine which architecture performs best for privacy prediction by investigating the performance of privacy prediction models based on visual semantic features extracted from all four architectures, AlexNet, GoogLeNet, VGG-16, and ResNet pre-trained on object data of ImageNet. We extract deep visual features: (1) fc_6 -A, fc_7 -A, fc_8 -A and “prob-A” from AlexNet, (2) $loss_3$ -G and “prob-G” from GoogLeNet, (3) fc_6 -V, fc_7 -V, fc_8 -V and “prob-V” from VGG-16, and (4) fc -R and “prob-R” from ResNet. For AlexNet and GoogLeNet, we used the pre-trained networks that come with the CAFFE open-source framework for CNNs⁴². For VGG-16, we used an improved version of pre-trained models presented by the VGG-16 team in the ILSVRC-2014 competition¹⁹. For ResNet, we use the ResNet pre-trained models of 101 layers given by He et al.²⁰.

Table 3.1 shows the performance (Accuracy, F1-measure, Precision, Recall) of SVMs trained on the features extracted from all four pre-trained networks. From the table, we can observe that the models trained on the features extracted from ResNet consistently yield the best performance. For example, ResNet achieves an F1-measure of 0.872 as compared with 0.849, 0.861, 0.864 achieved by AlexNet, GoogLeNet, and VGG-16, respectively. These results suggest that the deep Residual Networks have more representational abilities compared to the other networks, and are more effective for predicting appropriate privacy classes of images. Additionally, ResNets are substantially deeper than their “plain” counterparts, which allows extracting various image-specific features that are beneficial for learning images’ privacy characteristics better. Since privacy involves understanding the complicated relationship between the objects present in images, the features derived from ResNet prove to be more adequate than the features obtained by simply stacking convolutional layers. In

			Overall			Private			Public		
Features	H-Param	Acc %	F1	Prec	Re	F1	Prec	Re	F1	Prec	Re
AlexNet											
fc ₆ -A	R,1.0,0.05	82.29	0.82	0.819	0.823	0.613	0.639	0.591	0.885	0.875	0.895
fc ₇ -A	R,2.0,0.01	82.97	0.827	0.825	0.83	0.627	0.656	0.602	0.889	0.878	0.901
fc ₈ -A	R,1.0,0.05	<i>85.51</i>	<i>0.849</i>	<i>0.849</i>	<i>0.855</i>	<i>0.661</i>	<i>0.746</i>	<i>0.595</i>	<i>0.908</i>	<i>0.881</i>	<i>0.936</i>
prob-A	R,5.0,1.0	82.76	0.815	0.816	0.828	0.568	0.704	0.477	0.892	0.851	0.937
GoogLeNet											
loss ₃ -G	P,0.001,2.0	<i>86.42</i>	<i>0.861</i>	<i>0.86</i>	<i>0.864</i>	<i>0.695</i>	<i>0.746</i>	<i>0.652</i>	<i>0.913</i>	<i>0.895</i>	<i>0.93</i>
prob-G	R,50.0,0.05	82.66	0.815	0.816	0.827	0.573	0.694	0.488	0.891	0.853	0.933
VGG-16											
fc ₆ -V	R,1.0,0.01	83.85	0.837	0.836	0.839	0.652	0.67	0.636	0.895	0.888	0.902
fc ₇ -V	R,2.0,0.01	84.43	0.843	0.842	0.844	0.663	0.684	0.644	0.899	0.891	0.907
fc ₈ -V	R,2.0,0.05	<i>86.72</i>	<i>0.864</i>	<i>0.863</i>	<i>0.867</i>	<i>0.7</i>	<i>0.758</i>	<i>0.65</i>	<i>0.915</i>	<i>0.895</i>	<i>0.935</i>
prob-V	R,2.0,0.05	81.72	0.801	0.804	0.817	0.528	0.687	0.429	0.887	0.84	0.939
ResNet											
fc-R	R,1.0,0.05	87.58	0.872	0.872	0.876	0.717	0.783	0.662	0.92	0.899	0.943
prob-R	R,2.0,0.05	80.6	0.784	0.789	0.806	0.473	0.67	0.366	0.881	0.826	0.943

Table 3.1: Comparison of SVMs trained on features extracted from pre-trained architectures AlexNet, GoogLeNet, VGG-16 and ResNet. The best performance is shown in bold and blue color. The best performance for each network is shown in italics and orange color.

Table 3.1, we also show the class-specific privacy prediction performance in order to identify which features characterize the private class effectively as sharing private images on the Web with everyone is not desirable. Interestingly, we found that the model trained on features obtained from ResNet provides improved F1-measure, precision, and recall for the private class. Precisely, F1-measure for the private class improves from 0.661 (for AlexNet) to 0.717 (for ResNet), yielding an improvement of 6%. Similarly, for precision and recall, we obtain an increase of 4% and 7%, respectively, using ResNet features over the AlexNet features.

From Table 3.1, we also notice that the overall best performance (shown in orange and blue color) obtained for each network is higher than $\approx 85\%$ in terms of all compared measures (overall - Accuracy, F1-measure, precision and recall). Note that a naive baseline which classifies every image as “public” obtains an accuracy of 75%. Additionally, analyzing the results obtained by the VGG-16 features, we notice that as we ascend the fully-connected

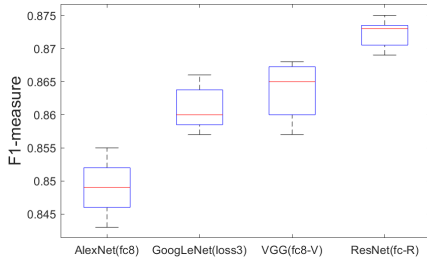


Figure 3.7: *Box plot of F1-measure (overall) obtained for the best-performing features derived from each CNN over five splits.*

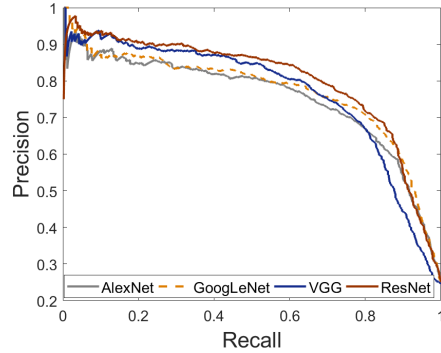


Figure 3.8: *Precision-recall curves for the private class obtained using features extracted from all four architectures AlexNet (fc_8), GoogLeNet ($loss_3$), VGG-16 (fc_8-V) and ResNet ($fc-R$).*

layers of the VGG-16 network from fc_6-V to fc_8-V , the F1-measure improves from 0.837 to 0.864 (see Table 3.1). Similarly, for AlexNet, the F1-measure improves from 0.82 (for fc_6-A) to 0.849 (for fc_8-A). This shows that the high-level object interpretations obtained through the last fully-connected layer helped to derive better privacy characteristics. Moreover, it is worth noting that “prob” features perform worse than the features extracted from the fully-connected layers (on all architectures). For example, prob-G obtains an F1-measure of 0.815, whereas $loss_3-G$ achieves an F1-measure of 0.861. One possible explanation could be that squashing the values at the previous layer (e.g., $loss_3-G$ in GoogLeNet) through the softmax function, which yields the “prob” layer, produces a non-linearity that is less useful for SVM compared to the untransformed values. We also experimented with a combination of features, e.g., fc_7-A concatenated with fc_8-A , but we did not obtain a significant improvement over the individual (fc_7-A and fc_8-A) features.

We also analyze the performance by showing the box plots of F1-measure in Figure 3.7, obtained for the most promising features of all the architectures over the five random splits of the dataset. The figure indicates that the model trained on ResNet features is statistically significantly better than the models that are trained on the features derived from the other architectures. We further compare features derived through all the architectures using precision-recall curves given in Figure 3.8. The curves show again that features derived

Features	H-Param	Acc %	Overall			Private			Public		
			F1	Prec	Re	F1	Prec	Re	F1	Prec	Re
Fine-tuned AlexNet											
ft-A	fc	85.01	0.846	0.845	0.851	0.657	0.723	0.606	0.904	0.883	0.926
ft-A	fc-all	85.14	0.849	0.847	0.852	0.669	0.713	0.632	0.904	0.889	0.92
ft-A	all	85.07	0.848	0.847	0.851	0.67	0.707	<i>0.638</i>	0.904	0.89	0.917
Pre-trained AlexNet											
fc ₈ -A	R,1,0.05	85.51	0.849	0.849	0.855	0.661	<i>0.746</i>	0.595	0.908	0.881	0.936
Fine-tuned GoogLeNet											
ft-G	fc	86.27	0.86	0.859	0.863	0.694	0.74	0.653	0.911	0.895	0.928
ft-G	all	86.77	0.867	0.867	0.868	<i>0.717</i>	0.732	<i>0.705</i>	0.914	0.909	0.919
Pre-trained GoogLeNet											
loss ₃ -G	P,0.001,2	86.42	0.861	0.86	0.864	0.695	0.746	0.652	0.913	0.895	0.930
Fine-tuned VGG-16											
ft-V	fc	86.74	0.864	0.865	0.869	0.695	<i>0.782</i>	0.631	0.916	0.891	<i>0.944</i>
ft-V	fc-all	86.92	0.869	0.87	0.869	0.722	0.73	0.717	0.914	0.912	0.917
ft-V	all	86.76	0.867	0.867	0.868	0.718	0.729	0.709	0.913	0.91	0.917
Pre-trained VGG-16											
fc ₈ -V	R,2,0.05	86.72	0.864	0.863	0.867	0.700	0.758	0.65	0.915	0.895	0.935
Fine-tuned ResNet											
ft-R	fc	87.23	0.87	0.869	0.873	0.717	0.759	<i>0.68</i>	0.918	0.903	0.932
ft-R	all	86.19	0.856	0.856	0.863	0.672	0.776	0.594	0.913	0.881	0.946
Pre-trained ResNet											
fc-R	R,1,0.05	87.58	0.872	0.872	0.876	0.717	0.783	0.662	0.92	0.899	0.943

Table 3.2: *Fine-tuned networks vs. Pre-trained networks. The best performance is shown in bold and blue color. The performance measures that achieve a better performance after fine-tuning a CNN over pre-trained features are shown in italics and orange color.*

from ResNet perform better than the features obtained from the other architectures, for a recall ranging from 0.5 to 0.8. For example, for a recall of 0.7, we achieve a precision of 0.75, 0.8, 0.8 and 0.85 for AlexNet, GoogLeNet, VGG-16, and ResNet, respectively.

3.6.3 Fine-Tuned Networks vs. Pre-Trained Networks

Previous works showed that the features transferred from the network pre-trained on the object dataset to the privacy data achieved a good performance¹⁴. Moreover, many other works used “transfer learning” to get more dataset specific features^{107;108}. Thus, we deter-

mine the performance of fine-tuned networks on the privacy dataset. We compare fine-tuned networks of all four architectures with the deep features obtained from pre-trained networks. We refer the fine-tuned networks of AlexNet, GoogLeNet, VGG-16, and ResNet as “ft-A,” “ft-G,” “ft-V,” and “ft-R” respectively. For fine-tuning, we used the same CNN architectures pre-trained on the object dataset, and employed in previous experiments. To fine-tune the networks, we experiment with the three types of settings: (1) fine-tune the last fully-connected layer (that has two output units corresponding to 2 privacy classes) with higher learning rates as compared to the learning rates of the rest of the layers of the networks (0.001 vs. 0.0001), referred as “fc.” (2) fine-tune all the fully-connected layers of the networks with higher learning rates and convolutional layers are learned with smaller learning rates. We refer to this setting as “fc-all.” (3) fine-tune all layers with the same learning rates and denoted as “all.” Note that since ResNet and GoogLeNet have only one fully-connected layer, we report the performance obtained only using “fc,” and “all” settings. The very low learning rate avoids substantial learning of the pre-trained layers. In other words, due to a very low learning rate (0.0001), pre-trained layers learn very slowly as compared to the layers that have a higher learning rate (0.001) to learn the required weights for privacy data.

Table 3.2 shows the performance comparison of the models obtained by fine-tuning architectures on privacy data and the models trained on the features derived from the pre-trained networks. We notice that we get mostly similar results when we fine-tune pre-trained models on our privacy dataset as compared to the models trained on the features derived from the pre-trained architectures. However, we get improved recall for the private class when we fine-tune the networks on the privacy dataset. For example, the fine-tuned VGG-16 network gets an improvement of 6.7% in the recall for the private class (see ft-V, fc-all setting vs. fc₈-V) over the models trained on the features extracted from the pre-trained VGG-16. The performance measures that achieve a better performance after fine-tuning a CNN over pre-trained features are shown in italics and orange color for each network. We notice that the fine-tuned VGG gives the best performance for the F1-measure and recall of the private class (shown in bold and blue color). However, the models trained on the features derived from the pre-trained ResNet yield the best overall performance (shown in bold and blue

color). Thus, we compare the models trained on fc-R features with prior privacy prediction approaches in the next subsection.

3.6.4 ResNet Features-Based Models vs. Prior Works

We compare the performance of the state-of-the-art works on privacy prediction, as detailed below, with the models trained using ResNet features, i.e., fc-R.

1. PCNH privacy framework¹⁴: This framework combines features obtained from two architectures: one that extracts convolutional features (size = 24, referred as Convolutional CNN), and another that extracts object features (size = 24, referred as Object CNN). The Convolutional CNN contains two convolutional layers and three fully-connected layers of size 512, 512, 24, respectively. On the other hand, the object CNN is an extension of AlexNet architecture that appends three fully-connected layers of size 512, 512, and 24, at the end of the last fully-connected layer of AlexNet and form a deep network of 11 layers. The two CNNs are connected at the output layer. The PCNH framework is first trained on the ImageNet dataset and then fine-tuned on a small privacy dataset.

2. AlexNet features^{24;27}: We consider the model trained on the features extracted from the last fully-connected layer of AlexNet, i.e., fc₈-A as another baseline, since in our previous works we achieved a good performance using these features for privacy prediction.

3. SIFT & GIST^{7;12;13}: We also consider classifiers trained on the best performing features between SIFT, GIST, and their combination as our baselines. Our choice of these features is motivated by their good performance over other visual features such as colors, patterns, and edge directions in prior works^{7;12}. For SIFT, we construct a vocabulary of 128 visual words for our experiments. We tried different numbers of visual words such as 500, 1000, etc., but we did not get a significant improvement over the 128 visual words. For a given image, GIST is computed by first convolving the image with 32 Gabor filters at 4 scale and 8 orientations, which produces 32 feature maps; second, dividing the feature map into a 4×4 grid and averaging feature values of each cell; and third, concatenating these 16 averaged values for 32 feature maps, which results in a feature vector of 512 (16×32)

			Overall			Private			Public		
Features	H-Param	Acc %	F1	Prec	Re	F1	Prec	Re	F1	Prec	Re
Highest performing CNN architecture											
fc-R	R,1.0,0.05	87.58	0.872	0.872	0.876	0.717	0.783	0.662	0.92	0.899	0.943
#1 PCNH framework											
PCNH	–	83.13	0.824	0.823	0.831	0.624	0.704	0.561	0.891	0.863	0.921
#2 AlexNet Deep Features											
fc _g -A	R,1.0,0.05	85.51	0.849	0.849	0.855	0.661	0.746	0.595	0.908	0.881	0.936
#3 SIFT & GIST models											
SIFT	P,1.0,2.0	77.31	0.674	0.598	0.773	0.002	0.058	0.001	0.87	0.772	0.995
GIST	R,0.001,0.5	77.33	0.674	0.598	0.773	0.002	0.058	0.001	0.87	0.772	0.995
SIFT & GIST	R,0.05,0.5	72.67	0.704	0.691	0.727	0.27	0.343	0.223	0.832	0.793	0.874
#4 Rule-based models											
Rule-1	–	77.35	0.683	0.694	0.672	0.509	0.47	0.556	0.853	0.875	0.832
Rule-2	–	77.93	0.673	0.704	0.644	0.458	0.373	0.593	0.897	0.914	0.88

Table 3.3: *Highest performing visual features (fc-R) vs. Prior works.*

length.

3. Rule-based classifiers: We also compare the performance of models trained on ResNet features fc-R with two rule-based classifiers which predict an image as *private* if it contains persons. Otherwise, the image is classified as *public*. For the first rule-based classifier, we detect front and profile faces by using Viola-Jones algorithm⁴³. For the second rule-based classifier, we consider user tags such as “women,” “men,” “people.” Recall that these tags are not present in the set of 1,000 categories of the ILSVRC-2012 subset of the ImageNet dataset, and hence, we restrict to user tags only. If an image contains one of these tags or detects a face, we consider it as “private,” otherwise “public.”

Table 3.3 compares the performance of models trained on fc-R features (the highest performing features obtained from our previous experiments) with the performance obtained by prior works. As can be seen from the table, the deep features extracted from the pre-trained ResNet achieve the highest performance, and hence, are able to learn the privacy characteristics better than the prior works with respect to both the classes. Precisely, using

fc-R features, F1-measure improves from 0.824 obtained by PCNH framework to 0.872 obtained by fc-R features, providing an overall improvement of 5%. Moreover, for the private class, fc-R features yield an improvement of 9.8% in F1-measure over the more sophisticated PCNH framework (from 0.624, PCNH to 0.717, fc-R features).

One possible explanation could be that the object CNN of PCNH framework is formed by appending more fully-connected layers to the AlexNet architecture and the increase in the number of complex non-linear layers (fully-connected layers) introduces more parameters to learn. At the same time, with a relatively small amount of training data (PicAlert vs. ImageNet), the object CNN model can over-fit. On the other hand, as images' privacy greatly depends on the objects in images, we believe that the low-level features controlling the distinct attributes of the objects (e.g., edges of swim-suit vs. short pants) obtained through the convolutional layers can better approximate the privacy function compared with adding more non-linear layers (as in PCNH). This is justified by the results, where the network with more convolutional layers, i.e., ResNet achieves a better performance as compared to the network with more fully-connected layers, i.e., PCNH. Additionally, even though PCNH attempted to capture convolutional features using Convolutional CNN, both CNN (convolutional and object) vary in their discriminative power and thus obtaining an optimal unification of convolutional CNN and object CNN is difficult. Moreover, PCNH is required to first train on ImageNet and then fine-tune on the PicAlert dataset. Training a deep network such as PCNH two times significantly increases the processing power and time. On the other hand, through our experiments, we found that the features derived from the state-of-the-art ResNet model can reduce the overhead of re-training and achieve a better performance for privacy prediction.

As discussed before, the models trained on ResNet features outperform those trained on AlexNet features. Interestingly, the best performing baseline among all corresponds to the SVM trained on the deep features extracted from the AlexNet architecture. For example, the SVM trained on the AlexNet features (fc₈-A) yields an F1-measure of 0.849 as compared with the F1-measure of 0.824 achieved by the PCNH framework. We hypothesize that this is due to the model complexity and the small size of the privacy dataset used to train the

PCNH framework. For example, merging two CNNs (as in PCNH) that vary in depth, width, and optimization algorithm can become very complex and thus the framework potentially has more local minima, that may not yield the best possible results. Additionally, unlike Tran et al.¹⁴, that used 800 images in their evaluation, we evaluate the models on a large set of images (32000), containing a large variety of image subjects. The features derived from the various layers of the state-of-the-art AlexNet reduce the overhead of training the complex structure and still achieve a good performance for privacy prediction.

Another interesting aspect to note is that, although we showed earlier that the fine-tuned network (in this case VGG-16) does not show a significant improvement over the ResNet pre-trained features (see Table 3.2), our fine-tuning approach yields better results compared to the PCNH framework. For example, fine-tuned VGG-16 (ft-V) achieves an F1-measure of 0.869 whereas PCNH achieves an F1-measure of 0.824 (see Tables 3.2 and 3.3). The possible reasons could be that we use a larger privacy dataset to fine-tune a simpler architecture, unlike PCNH that merges two convolutional neural networks. Additionally, we fine-tune the state-of-the-art VGG-16 model presented by Simonyan and Zisserman¹⁹, contrary to PCNH that required estimating optimal network parameters to train the merged architecture on the ImageNet dataset.

As expected, we can see from Table 3.3 that the baseline models trained on SIFT/GIST and the rule-based models are the lowest performing models. For example, the fc-R based models achieve improvement in performance as high as 17% over SIFT/GIST models. With a paired T-test, the improvements over the prior approaches for F1-measure are statistically significant for p-values < 0.05 . It is also interesting to note that rules based on facial features exhibit better performance than SIFT and GIST and suggest that feature representing persons are helpful to predict private images. However, fc-R features outperform the rule-based models based on facial features by more than 10% in terms of all measures.

We further analyze fc-R features and compare their performance with the prior works through precision-recall curves shown in Figure 3.9 (a). As can be seen from the figure, the SVM trained on ResNet features achieve a precision of ≈ 0.8 for recall values up to 0.8, and after that, the precision drops steadily.

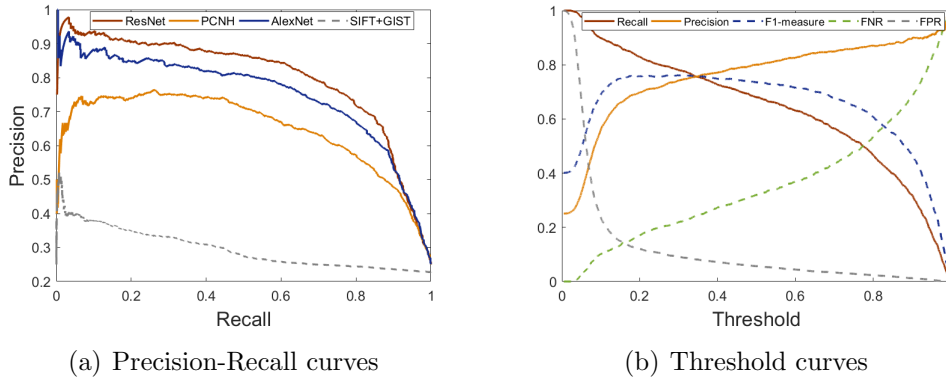


Figure 3.9: *Precision-Recall and Threshold curves for the private class obtained using ResNet features (fc-R) and prior works.*

The performance measures shown in previous experiments are calculated using a classification threshold of 0.5. In order to see how the performance measures vary for different classification thresholds, we plot the threshold curve and show this in Figure 3.9 (b). From the figure, we can see that the precision increases from ≈ 0.68 to ≈ 0.97 at a slower rate with the classification threshold. The recall slowly decreases to 0.8 for a threshold value of ≈ 0.4 , and the F1-measure remains comparatively constant until ≈ 0.75 . At a threshold of ≈ 0.4 , we get equal precision and recall of ≈ 0.78 , which corresponds to the breakeven point. In the figure, we also show the false negative rate and false positive rate, so that depending on a user’s need (high precision or high recall), the classifier can run at the desired threshold. Also, to reduce the number of content-sensitive images shared with everyone on the Web, lower false negative (FN) rates are desired. From Figure 3.9 (b), we can see that we achieve lower FN rates up to ≈ 0.4 for the threshold values up to 0.8.

3.6.5 Best Performing Visual Features vs. Tag Features

Image tags provide relevant cues for privacy-aware image retrieval⁷ and can become an essential tool for surfacing the hidden content of the deep Web without exposing sensitive details. Additionally, previous works showed that user tags performed better or on par compared with visual features^{7;12;24;27}. For example, in our previous work^{24;27}, we showed that the combination of user tags and deep tags derived from AlexNet performs comparably to

			Overall			Private			Public		
Features	H-Param	Acc %	F1	Prec	Re	F1	Prec	Re	F1	Prec	Re
Best performing CNN architecture											
fc-R	R,1.0,0.05	87.58	0.872	0.872	0.876	0.717	0.783	0.662	0.92	0.899	0.943
#1 User Tags (BoT)											
UT	R,2.0,0.05	78.63	0.777	0.772	0.786	0.496	0.565	0.442	0.865	0.837	0.894
#2 Deep Tags (BoT)											
DT-A	R,1.0,0.1	83.34	0.825	0.824	0.833	0.601	0.699	0.529	0.895	0.863	0.929
DT-G	R,1.0,0.05	83.59	0.828	0.827	0.836	0.606	0.699	0.534	0.896	0.866	0.929
DT-V	P,1.0,1.0	83.42	0.826	0.825	0.834	0.607	0.698	0.537	0.895	0.865	0.927
DT-R	P,1.0,1.0	83.78	0.833	0.831	0.838	0.631	0.688	0.584	0.896	0.876	0.917
#3 User Tags & Deep Tags											
UT+DT-R (BoT)	R,1.0,0.05	84.33	0.84	0.839	0.843	0.67	0.709	0.636	0.897	0.882	0.913
Tag CNN	—	85.13	0.855	0.855	0.854	0.706	0.700	0.712	0.901	0.903	0.898

Table 3.4: *Visual features vs. Tag features.*

the AlexNet based visual features. Hence, in this experiment, we compare the performance of fc-R features with the tag features. For deep tags, we follow the same approach as in our previous work^{24;27} and consider the top $k = 10$ object labels since $k = 10$ worked best. “DT-A,” “DT-G,” “DT-V,” and “DT-R” denote deep tags generated by AlexNet, GoogLeNet, VGG-16, and ResNet, respectively. Deep tags are generated using the probability distribution over 1,000 object categories for the input image obtained by applying the softmax function over the last fully-connected layer of the respective CNN.

Table 3.4 compares the performance obtained using models trained on fc-R features with the performance of models trained on the tag features. We consider tag features as: (1) user tags (UT); (2) deep tags (DT) obtained from all architectures; (3) the combination of user tags and best performing deep tag features using Bag-of-Tags (BoT) model; and (4) Tag CNN applied to the combination of user and deep tags. As can be seen from the table, the visual features extracted from ResNet outperform the user tags and deep tags independently as well as their combination. The models trained on fc-R features achieve an improvement of 2% over the CNN trained on the combination of user tags and deep tags

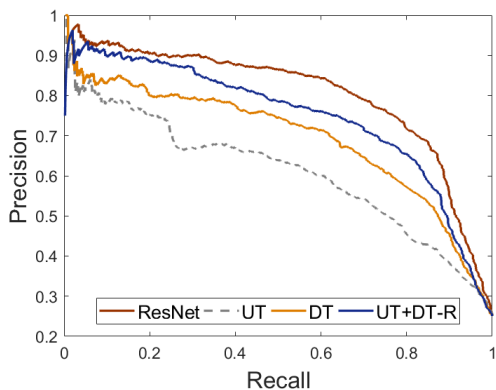


Figure 3.10: Precision-Recall curves for the private class obtained using visual features (fc -R) and tag features as user tags (UT), deep tags (DT-R), the combination of user tags and deep tags (UT + DT-R).

(Tag CNN). Additionally, the models trained on fc -R features yield an increase of 9.5% in the F1-measure over the user tags alone and an increase of 4% over the best performing deep tags, i.e., DT-R (among the deep tags of the four architectures).

From Table 3.4, we also observe that the Tag CNN performs better than the Bag-of-Tags model (DT-R+UT), yielding an improvement of 3.0% in the F1-measure of private class. Additionally, even though the visual features (fc -R) yield overall a better performance than the tag features, for the private class, the F1-measure (0.717) of the visual features (fc -R) is comparable to the F1-measure (0.706) of the Tag CNN. It is also interesting to note that the Visual CNN (fc -R) achieves an increase of 8% in the precision (private class) over the Tag CNN whereas the Tag CNN obtains an improved recall (private class) of 5% over the Visual CNN.

In order to see how precision varies for different recall values, we also show the precision-recall curves for the visual and tag features in Figure 3.10. To avoid clutter we show the precision-recall curves for deep tags derived through ResNet and the combination of user tags and deep tags (DT-R) using BoT model. From the curves, we can see that the ResNet visual features perform better than the tag features, for a wide range of recall values from 0.3 to 0.8.

We further analyze both the type of image encodings (visual & tag) by examining the privacy predictions obtained for anecdotal examples using both the encodings.

						
Features	(a)	(b)	(c)	(d)	(e)	(f)
Visual	<i>private</i>	<i>private</i>	<i>public</i>	<i>public</i>	<i>public</i>	<i>private</i>
Tags	<i>private</i>	<i>private</i>	<i>private</i>	<i>private</i>	<i>private</i>	<i>public</i>

Figure 3.11: *Privacy predictions obtained by image content encodings.*

Anecdotal Examples:

In order to understand the quality of predictions obtained by visual and tag features, we show privacy predictions for some samples obtained by both type of features. Figure 3.11 shows the predictions obtained using SVM models trained on the visual features and those trained on the combination of user tags and deep tags. Correct predictions are shown in italics and green in color. We can see that for images (a) and (b), the models trained on image tags (UT+DT) and visual features provide correct predictions. The tags such as “groom,” “bride,” “wedding,” “photography” describe the picture (a) adequately, and hence, using these tags appropriate predictions are obtained. Similarly, visual features identify the required objects, and a relationship between the objects and provide an accurate prediction for these images. Consider now examples (c) and (d). For these images, visual features capture the required objects to make accurate predictions, whereas, image tags such as “bruins,” “fight,” of image (c) and “cute,” “wool,” “bonnet” of image (d) do not provide adequate information about the picture and hence, yield an incorrect prediction. However, tags such as “hockey,” “sports” for image (c) and “toy,” “doll” for image (d) would have helped to make an appropriate prediction. We also show some examples, (e) and (f), for which visual features fail to predict correct privacy classes. Particularly, for image (f), we notice that visual features capture object information that identifies the image as private. On the other hand, the image tags such as “festival” and “sport” (describing the scene) provide additional information (over the object information) that helps the tag-based classifier to identify the picture as public.

Next, we provide the detailed analysis of image tags with respect to privacy.

Analysis of Image Tags with Respect to Privacy Classes:

Rank 1-10	Rank 11-20	Rank 21-30	Rank 31-40	Rank 41-50
people	pyjama	maillot	promontory	jersey
wig	jammies	girl	t-shirt	mole
portrait	sweatshirt	suit of clothes	foreland	groin
bow-tie	outdoor	ice lolly	headland	bulwark
neck brace	lakeside	suit	bandeau	seawall
groom	lakeshore	lollipop	miniskirt	seacoast
bridegroom	sun blocker	two-piece	breakwater	indoor
laboratory coat	sunscreen	tank suit	vale	stethoscope
hair spray	sunglasses	bikini	hand blower	valley
shower cap	military uniform	swimming cap	jetty	head

Table 3.5: Top 50 highly informative tags. We use the combination of deep tags and user tags (DT+UT) to calculate the information gain. User tags are shown in bold.



Figure 3.12: High frequency tag clouds with respect to public and private images.

We provide an analysis of the deep tags (capturing the visual content of the image) and user tags to learn their correlation with the private and public classes. First, we rank user tags and deep tags based on their information gain on the train set. Table 3.5 shows top 50 tags with high information gain. From the table, we observe that the tags such as “maillot,” “two-piece,” “sandbar” provide high correlation to the privacy classes. We also notice that deep tags (objects) contribute to a significant section of top 50 highly informative tags.

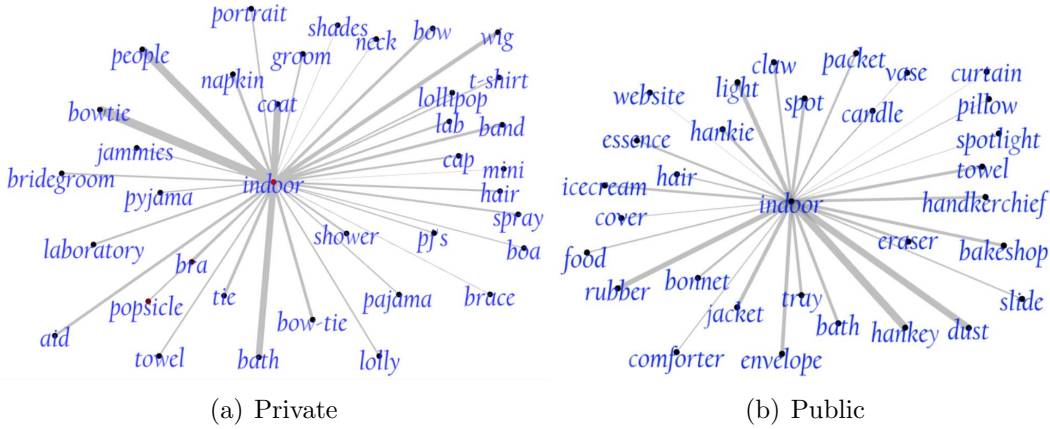


Figure 3.13: *Tag association graph.*

Secondly, we rank both the tags (user and deep tags) based on their frequency in public and private classes. We show 50 most frequent tags for each privacy class using word clouds in Figure 3.12. The tags with larger word size depict a higher frequency of the tag. We notice that tags such as “indoor,” “people,” “portrait” occur more frequently in the private class, whereas tags such as “outdoor,” “lakeside,” “fountain,” occur more frequently in the public class.

We also observe that some informative tags overlap in both public and private clouds (See Figure 3.12, e.g., “indoor”). Thus, we analyze other tags that co-occur with the overlapping tags to further discriminate between their association with the public and private classes. To inspect the overlapping tags, we create two graphs with respect to public and private classes. For the public graph, we consider each tag as a node in the graph and draw an edge between the two nodes if both the tags belong to the same public image. Likewise, we construct another graph using private images. Figure 3.13 shows portions of both public and private graphs for “indoor” tag. To reduce the complexity of visualization, we only display nodes with stronger edges that have the co-occurrence greater than a certain threshold. Note that stronger edges (edges with higher width) represent the high co-occurrence coefficient between two nodes (in our case, tags). From the graphs, we observe that the overlapping tag “indoor” tends to have different highly co-occurring tags for public and private classes. For example, the “indoor” tag shows high co-occurrence with tags such as “people,” “bath,”

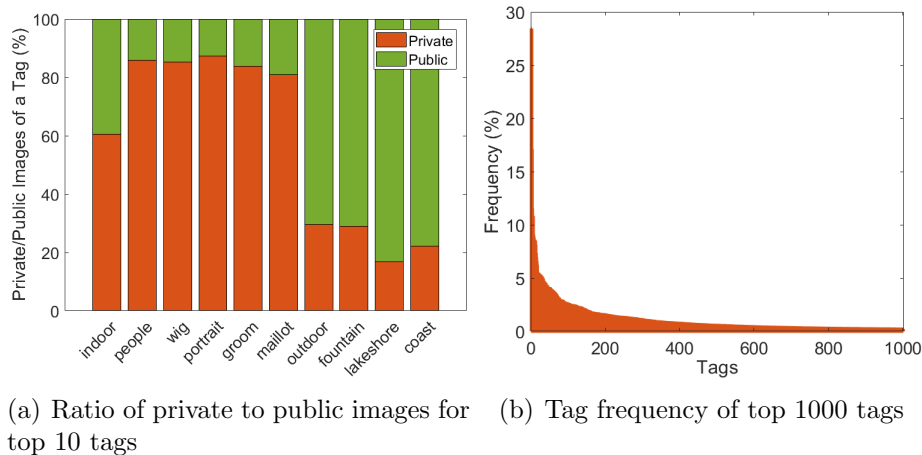


Figure 3.14: *Analysis of top frequently occurring tags.*

“coat,” “bowtie,” “bra” (tags describing private class) in the private graph. On the other hand, in the public graph, the tag shows high co-occurrence with “dust,” “light,” “hankey,” “bakeshop,” “rubber,” and so forth (tags describing public class). Even though some tags in the graph have comparatively low co-occurrence, the tags occurring in the private graph tend to associate with the private class whereas the tags from the public graph are more inclined towards the public class.

We further analyze the privacy differences of top 10 private and public image subjects. We consider “outdoor,” “indoor,” “fountain,” “lakeshore,” and “coast” for the public class. On the other hand, we consider “indoor,” “people,” “wig,” “portrait,” “outdoor,” “groom,” and “maillot” for the private class. Note that since images may have various tags associated with them, an image can be counted towards more than one tag. Given that the dataset contains three times more public images than private images (3 : 1 public to private ratio), we count 3 for each private image as opposed to the public class where we count 1 for each public image for a fair comparison. The ratio of private to public content for a specific tag is shown in Figure 3.14 (a). For example, out of the total images that possess the “indoor” tag, 60% images are of private class. From the figure, we observe that tags except for “indoor” show a significant difference in the inclination towards public and private classes. We also plot the frequency of top 1000 tags normalized by the dataset size in Figure 3.14 (b). The plot shows that the top 200 tags befall in 3% – 30% of the dataset with very few tags

Features	Overall				Private			Public		
	Acc %	F1	Prec	Re	F1	Prec	Re	F1	Prec	Re
fc-R	87.58	0.872	0.872	0.876	0.717	0.783	0.662	0.92	0.899	0.943
fc-R+UT	88.29	0.881	0.88	0.883	0.753	0.799	0.713	0.923	0.907	0.94

Table 3.6: *Results for the combination of visual and tag features.*

occurring in around 20% of the dataset. We also observe that most of the tags lie below 3% of the dataset showing the variation in the images’ subjects and complexity of the dataset which justifies the fact that increasing the number of images increases the challenges of the problem statement.

3.6.6 Fusion of Visual and Tag Features for Image Privacy Prediction

Visual encoding and tag encoding capture different aspects of images. Thus, we add the top 350 correlated tags to the visual features fc-R and evaluate their performance for privacy prediction. We experiment with the number of top correlated tags = {10, 20, \dots , 50, 100, \dots , 500, 1000, 5000, 10000}. However, we get the best results with the top 350 correlated tags. Table 3.6 shows the results obtained using SVMs trained on fc-R and the combination of fc-R with the top 350 correlated user tags (fc-R+tag). The results reveal that adding the highly correlated tags improves the privacy prediction performance. Precisely, we get a significant improvement of 4% on F1-measure of private class over the performance obtained using visual features fc-R. Note that, in our previous works^{24;27} and Experiment 3.6.5 (where we compare visual and tag features), we described visual content using tags (deep tags) and combined with the user tag to achieve a better performance. However, the combination of user tags and deep tags (combining one type of encoding) yields a lower performance as compared to the combination of user tags and fc-R features (combining two types of encodings). Precisely, the combination of user tags (UT) and fc-R features yields an improvement of 5% in the F1-measure of private class (refer Tables 3.4 and 3.6) over the combination of user tags and deep tags.

3.7 Chapter Summary and Future Directions

In this chapter, we provide a comprehensive study of the deep features derived from various CNN architectures of increasing depth to discover the best features that can provide an accurate privacy prediction for online images. Specifically, we explored features obtained from various layers of the pre-trained CNNs such as AlexNet, GoogLeNet, VGG-16, and ResNet and used them with SVM classifiers to predict an image’s privacy as *private* or *public*. We also fine-tuned these architectures on a privacy dataset. The study reveals that the SVM models trained on features derived from ResNet perform better than the models trained on the features derived from AlexNet, GoogLeNet, and VGG-16. We found that the overall performance obtained using models trained on the features derived through pre-trained networks is comparable to the fine-tuned architectures. However, fine-tuned networks provide improved performance for the private class as compared to the models trained on pre-trained features. The results show remarkable improvements in the performance of image privacy prediction as compared to the models trained on CNN-based and traditional baseline features. Additionally, models trained on the deep features outperform rule-based models that classify images as private if they contain people. We also investigate the combination of user tags and deep tags derived from CNN architectures in two settings: (1) using SVM on the bag-of-tags features; and (2) applying the text CNN over these tags. We thoroughly compare these models with the models trained on the highest performing visual features obtained for privacy prediction. We further provide a detailed analysis of tags that gives insights for the most informative tags for privacy predictions. We finally show that the combination of deep visual features with these informative tags yields improvement in the performance over the individual sets of features (visual and tag).

The result of our classification task is expected to aid other very practical applications. For example, a law enforcement agent who needs to review digital evidence on a suspected equipment to detect sensitive content in images and videos, e.g., child pornography. The learning models developed here can be used to filter or narrow down the number of images and videos having sensitive or private content before other more sophisticated approaches can

be applied to the data. Consider another example, images today are often stored in the cloud (e.g., Dropbox or iCloud) as a form of file backup to prevent their loss from physical damages and they are vulnerable to unwanted exposure when the storage provider is compromised. Our work can alert users before uploading their private (or sensitive) images to the cloud systems to control the amount of personal information (eg. social security number) shared through images.

In the future, using this study, an architecture can be developed, that will incorporate other contextual information about images such as personal information about the image owner, owner's privacy preferences or the owner social network activities, in addition to the visual content of the image. Another interesting direction is to extend these CNN architectures to describe and localize the sensitive content in private images.

Chapter 4

Privacy-Aware Tag Recommendation for Image Sharing

Online images' tags are very important for indexing, sharing, and searching of images, as well as surfacing images with private or sensitive content, which needs to be protected. Social media sites such as Flickr generate these metadata from user-contributed tags. However, as the tags are at the sole discretion of users, these tags tend to be noisy and incomplete. In this chapter, we present a privacy-aware approach to automatic image tagging, which aims at improving the quality of user annotations, while also preserving the images' original privacy sharing patterns. Precisely, we recommend potential tags for each target image by mining privacy-aware tags from the most similar images of the target image, which are obtained from a large collection. Experimental results show that, although the user-input tags comprise noise, our privacy-aware approach is able to predict accurate tags that can improve the performance of a downstream application on image privacy prediction, and outperforms an existing privacy-oblivious approach to image tagging. The results also show that, even for images that do not have any user tags, our proposed approach can recommend accurate tags. Crowd-sourcing the predicted tags exhibits the quality of our privacy-aware recommended tags. Our code, features, and the dataset used in experiments are available at: <https://github.com/ashwinitonge/privacy-aware-tag-rec.git>.

4.1 Introduction

Images are constantly shared on social networking sites such as Facebook, Flickr, and Instagram. For instance, it is common to take photos at cocktail parties and upload them on social networking sites without much hesitation for self-promotion and personal sharing. However, when privacy settings are used inappropriately, these photos can potentially reveal a user’s personal and social habits, resulting in unwanted disclosure and privacy violations^{5;7;12;13;40}. For example, malicious attackers can take advantage of these accidental leaks to launch context-aware or even impersonation attacks. Personal data can be harvested through social media without users’ consent if the privacy settings of social media are not managed properly, which could lead to online privacy risks¹¹⁷. A study carried out by the Pew Research center reports that 11% of the users of social networking sites regret the content they posted¹⁰. Thus, several works have been developed in recent years in an attempt to provide appropriate privacy settings for online images^{7;12;14;24;26;27;32;40;88}.

Prior works on privacy prediction^{7;12;24;26;27;88} found that the tags associated with images are indicative of their sensitive content. Tags are also important for image-related applications such as indexing, sharing, searching, content detection and social discovery^{109–112}. Yet, the tags are at the sole discretion of users, and hence, they tend to be noisy and incomplete²³. Despite that many approaches to automatic image tagging have been developed^{118–123}, these approaches do not consider the privacy aspect of an image while making the annotations (or tagging) and could not be sufficient for identifying images’ private content.

We posit that visually similar images can possess very different sets of tags if these images have different privacy orientations. For example, Figure 4.1 shows anecdotal evidence obtained from a Flickr dataset in which visually similar images of private and public classes display different sets of user tags. The picture of a woman that belongs to the private class in Figure 4.1(a) contains tags such as “Elegant,” “Corporate,” “Style,” and “Pretty,” whereas the picture of a woman that belongs to the public class in Figure 4.1(b) contains tags such as “Celebrity,” “Famous,” “News,” and “Hollywood.” An image is considered to be private if it belongs to the private sphere (e.g., portraits, family, friends, home) or contains information



- (a) *Private*: Stylish, Elegant
Corporate, Style, Street
Fashion, Girl, Woman
Skirt, Top, Bag, Pretty
- (b) *Public*: Parisi, Sabrina
News, Celebrity
Famous, Girl
Woman, Hollywood

Figure 4.1: Anecdotal evidence for visually similar images with privacy-aware user tags.

that can not be shared with everybody on the Web (e.g., private documents), whereas the remaining images are considered to be public⁷. Figure 4.1 shows that the images’ tags are correlated to each image’s privacy patterns^{6;68;88}. These tags are very useful when access to the visual content of images is not allowed due to users reluctance to share the actual images for visual content analysis (which could reveal a user’s identity through the face and friends, etc.). In such cases, privacy-aware tags can become good indicators of the privacy settings and can help improve the privacy prediction methods to reduce privacy breaches.

To this end, we ask the following questions: *Can we develop an automated approach to recommend accurate image tags that can also take into account the sharing needs of the users for images in questions? Can this method make precise tag recommendations for newly uploaded images that have an incomplete set of user tags or no tags at all? Can these recommended tags help improve the image privacy prediction performance?* We address these questions with our research agenda. In particular, we draw ideas from the collaborative filtering line of research and explore its applicability to privacy-aware image tagging. Collaborative filtering is widely used to make recommendations for unknown items to users and relies on the assumption that similar users express similar interests or preferences on similar items¹²⁴. Hence, we explore tag recommendation to images based on images’ similar neighbors.

Our contributions are as follows:

- We present a privacy-aware approach to automatic image tagging, that aims at improving the quality of user annotations (or user tags), while also preserving the images' original privacy sharing patterns. Precisely, we recommend potential tags for each target image by mining privacy-aware tags from the most similar images of the target image, which we obtain from a large collection of images.
- We study our privacy-aware recommended tags obtained by the proposed privacy-aware weighting scheme in an ablation experiment for privacy prediction. In this experiment, we compare various privacy-aware and privacy-oblivious weighting schemes and observe how the privacy prediction performance varies for these weighting schemes. We also experiment with various parameter values to estimate the best parameter setting.
- We compare the performance of privacy prediction using tags recommended by the proposed approach against the tags recommended by a prior state-of-the-art image annotation method. Our objective in this experiment is to verify whether the recommended tags by the proposed approach can capture better privacy characteristics than the prior state-of-the-art annotation.
- We investigate tag recommendation in a binary image privacy prediction task and show that the predicted tags can exhibit relevant cues for specific privacy settings (*public* or *private*) that can be used to improve the image privacy prediction performance.
- Our results show that we achieve a better privacy prediction performance when we add the recommended privacy-aware tags to the original user tags and predicted deep tags of images as compared to prior approaches of image privacy prediction.
- We also evaluate the recommended tags by employing crowd-sourcing to identify relevancy of the suggested tags to images. The results show that, although the user-input tags comprise noise or even some images do not have any tags at all, our approach is able to recommend accurate tags. In addition, we evaluate both privacy-aware and privacy-oblivious recommended tags and show that the privacy-aware recommended

tags describe an image’s content more accurately as compared to the privacy-oblivious tags.

4.2 Related work

In this section, we briefly review the related work on three lines of research: 1) automatic image annotation, 2) tag recommendation using collaborative filtering, and 3) online image privacy.

4.2.1 Automatic Image Annotation

Numerous approaches to automatic image annotation (or tagging) have been proposed in the literature to improve the search and retrieval of images based on text queries. We classify these methods into five categories: Generative methods that maximize the generative likelihood of image features and tags; Discriminative methods that consider the image annotation task as a multi-label classification problem; Tag completion methods that predict tags by automatically identifying the missing labels and also correcting the noisy labels for images; Deep learning based methods that train deep networks for predicting labels; Nearest neighbors methods that obtain tags from the similar images of the target image.

Generative methods.

The generative methods try to maximize the generative likelihood of the image features and tags^{118;125–139}. For example, Lavrenko et al.¹²⁵ learned joint probabilistic models of image content features and tags. These models compute the conditional likelihood of words given image content features that can be used to infer the most likely tags for an image. Similarly, Feng et al.¹¹⁸ learned a joint probability distribution of tags and image features in a relevance model framework. Other approaches treated the problem as statistical inference in graphical models, e.g., Latent Dirichlet Allocation (LDA) like models^{126–128}. Feng and Lapata¹²⁹ proposed an approach to automatically annotate images embedded in news articles by

using words from captions and article’s content. They used the idea that the accompanying document can provide accurate tags and reduce the effect of noise comprised in captions. Later, Feng and Lapata¹³⁰ used LDA to infer topics that capture co-occurrences of visual features and words.

Discriminative methods.

Discriminative methods perceive image annotation as a multi-label classification problem. In these works, the authors typically treat the image tagging as a classification task and train classifiers (e.g., Support Vector Machines) for each tag using image’s textual and/or visual features^{140–145}. In recent years, the graph-based learning (semi-supervised) methods are also used for image annotation in which the model is the graph of the entire data. The label correlation is incorporated in the graph as graph weights^{120;146–150} or as an additional constraint^{151;152}. In addition to the graph-based learning methods, some studies exploit the local label correlations¹⁵³, underlying correlations among labels using a multi-label dictionary learning algorithm¹⁵⁴ and handle the missing tags issues¹⁵⁵. On the other hand, Li et al.¹⁵⁶ considered image annotation as multi-correlation learning to the rank problem where the visual similarity among images and the semantic relevance among tags are explored simultaneously. Ivasic-Kos et al.¹⁵⁷ proposed a two-tier annotation model wherein the first tier corresponds to object-level annotation and the second tier to scene-level (image context) annotation.

Tag completion methods.

The tag completion methods automatically annotate images by identifying the missing tags and correcting the noisy tags. The entire dataset is represented as an initial matrix with each row as an image and each column as a tag. The tag completion methods recover this initial matrix by identifying correct associations between images and labels. The tag completion-based annotation is achieved by matrix completion^{158;159}, linear sparse reconstruction^{160;161}, subspace clustering with matrix completion¹⁶², and low-rank matrix factorization^{163;164}.

Deep learning methods.

The deep-learning based image annotation adopt image features and semantic tag relationships extracted using deep networks^{165–167}. For example, Gong et al.¹⁶⁷ used weighted approximate ranking to train deep convolutional neural networks (CNN) for multi-label image annotation tasks. Wang et al.¹⁶⁸ proposed a multitask voting automatic image annotation CNN, which contains shallow layers and regards each category as a label directly, using the raw images as inputs for large scale image annotation. Yang et al.¹⁶⁹ provided a multi-view stacked auto-encoder model for image annotation. Wang et al.¹⁷⁰ proposed a framework in which CNN and recurrent neural network (RNN) are jointly utilized to derive an image representation and the correlation between the adjacent labels, based on which the probabilities of the labels are computed. Jin and Nakayama¹⁷¹ also used the CNN-RNN framework for image annotation and Wu et al.¹⁷² proposed a joint deep multi-instance learning framework that learns objects and images’ tags simultaneously.

Nearest neighbors methods.

The nearest neighbor model-based image annotation methods assume that visually similar images are more likely to share common labels^{173–177}. For a given target image, these methods first obtain a set of similar images and then the tags of the target image are derived based on the tags of the similar images. For example, Makadia et al.¹²¹ proposed a joint equal contribution model that utilizes global low-level image features and a simple combination of basic distance measures to find nearest neighbors of a given image. Tags are then assigned from the nearest neighbor(s) based on the co-occurrence of tags. Guillaumin et al.¹¹⁹ proposed the “TagProp” model, which integrates a weighted nearest neighbor based method and metric learning capabilities into a discriminative framework. Chen et al.¹²² proposed an approach called “FastTag” to image annotation, in which the authors learned two classifiers to predict tags: one that reconstructs the complete tag set from the tags available during training and the other that maps image features to the reconstructed tag set. Kalayeh et al.¹⁷⁸ developed a hybrid model by integrating the nearest-neighbor scheme into

a generative model.

Furthermore, Cheng et al.¹⁷⁹ discussed advantages and disadvantages of these methods in details. For instance, the generative models may not be able to capture the intricate relationship between image features and labels, which is imperative to identify the privacy-aware tags. Additionally, the multi-label classification based discriminative approaches cannot be extended to a large number of image tags since a binary classifiers has to be trained for each tag, which is not feasible for the online images that contain diverse sets of tags. The authors also discussed that the correlation between image visual features and labels are often neglected by some discriminative methods. On the other hand, the tag completion models suffer from a major weakness, that is, the transformation of the tag completion process to an optimization problem. The process of optimizing the objective function may be time-consuming and computationally very complex, and cannot guarantee global optimization. Moreover, despite that deep learning based methods have shown significant improvements in the performance of image annotation, there are still a few shortcomings with these methods. The main drawback is that although RNN + CNN solve issues pertaining to label quantity prediction and label dependencies for large-scale image annotation, still a better solution to rank labels is needed as RNN requires an ordered sequential list as input, which is mostly not present in the online images. Another drawback is that the increase in the depth and breadth of the deep networks can cause the decrease in the efficiency of annotation methods. The nearest neighbor based methods are clear and intuitive, and many of them have been proven to be quite successful for tag prediction due to their high flexibility. However, improvements are still needed because of some inherent shortages. For instance, the performance of these methods is highly sensitive to the retrieval performance. Thus, an efficient way to identify appropriate neighbors for unlabeled images is highly sought.

In contrast to previous annotation mechanisms, we take advantage of both nearest neighbors and deep learning based approaches to provide privacy-aware image annotations. We consider nearest neighbor based approaches as our strong baselines.

4.2.2 Tag Recommendation using Collaborative Filtering

Our tag recommendation approach draws ideas from collaborative filtering, and hence, here we briefly review the most relevant works on tag recommendation using collaborative filtering. Xu et al.¹⁸⁰ designed a collaborative filtering approach to suggest high-quality tags for Web objects, according to several criteria (coverage, popularity, effort, uniformity). The authors employed a co-occurring strategy and considered that if two tags frequently co-occur when describing a specific object, they should also co-occur in the recommended set of tags. A similar approach was presented later by Sigurbjörnsson and van Zwol¹⁸¹, who recommended tags for Flickr images. They used knowledge from the Flickr community and applied it in a co-occurring strategy. Specifically, given a user-input tag, they considered the tags co-occurring with it as good candidates for recommendation. Peng et al.¹⁸² designed a novel technique for collaborative filtering in social tagging systems, in which all the interactions among users, items and tags are leveraged. They generated joint item-tag recommendations for users, where the tags represent topics from an item (i.e., a web resource) in which the user may be interested. Seitlinger et al.¹⁸³ used a model of human category learning (i.e., ALCOVE) for social tags recommendation. The model uses semantic information regarding a user-specific bookmark (e.g., Wikipedia categories or LDA topics). Tags are predicted to a user by applying the semantic information to a connectionist network with three layers, which simulates the user’s categorization and the bookmark formalization.

Recently, several works have been proposed to recommend tags for visual content types^{184–189}. For example, Liu et al.¹⁸⁴ explored locations to recommend tags to images.¹⁸⁵ proposed a system to automatically recommend tags to YouTube videos based on their audio-visual content. Gong and Zhang¹⁸⁶ adopted CNNs to recommend hashtags for microblogs. Zhang et al.¹⁸⁷ proposed a co-attention network incorporating textual and visual information to recommend hashtags for multimodal tweets. Nguyen et al.¹⁸⁸ presented a personalized content-aware image tag recommendation approach that combines both historical tagging information and image-based features in a factorization model. Seah et al.¹⁸⁹ concurrently generated ranked lists of comments and tags of a social image based on their joint relevance to the visual

features, user comments, and user tags.

In contrast to these works, we recommend privacy-aware tags for images shared online.

4.2.3 Online Image Privacy

The rapid increase in images uploaded on the Web intrigued researchers to focus on establishing adequate privacy models to help protect users' sensitive information. Researchers also provided public awareness of privacy risks associated with images shared online^{44;45}. Along this line, several works were carried out to study users' privacy concerns in social networks, privacy decisions about sharing resources, and the risk associated with them^{8;46-51;190}. Additionally, several works on privacy analysis examined privacy decisions and considerations in mobile and online photo sharing^{5;8;52;53}. For example, Ahern et al.⁵ studied the effectiveness of location information and tags in predicting privacy settings of images. They also conducted a study to verify whether the visual features are relevant to an image's privacy and found that content is one of the discriminatory factors affecting image privacy, especially for images depicting people. This supports the core idea underlying our work: that tags depicting private categories obtained from image content are pivotal for identifying the sensitive content from the search results. For example, tags such as "wedding," "bride," "people" describing a wedding event (private category) represent the private class that particular categories of image content are pivotal for identifying the sensitive content from the search results in establishing users' images sharing decisions. Jones and O'Neill⁵² reinforced the role of privacy-relevant image concepts. For instance, they determined that people are more reluctant to share photos capturing social relationships than photos taken for functional purposes; certain settings such as work, bars, concerts cause users to share less. Besmer and Lipford⁵³ discussed that users want to regain control over their shared content, but meanwhile, they feel that configuring proper privacy settings for each image is a burden.

In the context of automated approaches, image privacy is explored along four lines of research: *social group based approaches*, in which users' profiles are used to partition the

friends' lists into multiple groups or circles, and the friends from the same circle are assumed to share similar privacy preferences; *location-based approaches*, in which location contexts are used to control the location-based privacy disclosures; *visual-based approaches*, in which the visual content of images is leveraged for privacy prediction; and *tag-based approaches*, in which tags are used for privacy setting recommendations.

Social group based approaches.

Several works emerged to provide the automated privacy decisions for images shared online based on the social groups or circles^{7;54–69}. For example, Christin et al.⁵⁵ proposed an approach to share content with the users within privacy bubbles. Privacy bubbles represent the private sphere of the users and the access to the content is provided by the bubble creator to people within the bubble. Bonneau et al.⁶⁰ introduced the notion of privacy suites which recommend users a set of privacy settings that “expert” users or the trusted friends have already established so that ordinary users can either directly accept a setting or perform minor modifications only. Fang and LeFevre⁶¹ developed a privacy assistant to help users grant privileges to their friends. The approach takes as input the privacy preferences for the selected friends and then, using these labels, constructs a classifier to assign privacy labels to the rest of the (unlabeled) friends based on their profiles. Danezis⁶³ generated privacy settings based on the policy that the information produced within a social circle should remain in that circle itself. Along these lines, Adu-Oppong et al.⁷⁰ obtained privacy settings by forming clusters of friends by partitioning a user’s friends’ list. Yuan et al.⁶⁹ proposed an approach for context-dependent and privacy-aware photo sharing. This approach uses the semantics of the photo and the requester’s contextual information in order to define whether an access to the photo will be granted or not at a certain context. These social group based approaches mostly consider the user trustworthiness, but ignore the image content sensitiveness, and thus, they may not necessarily provide appropriate privacy settings for online images as the privacy preferences might change according to sensitiveness of the image content.

Location-based approaches.

These approaches^{69;71-81} leverage geo-tags, visual landmarks and other location contexts to control the location-based privacy disclosures. The geo-tags can be provided manually via social tagging or by adding location information automatically through the digital cameras or smart-phones having GPS. The location can also be inferred by identifying places from the shared images through the computer vision techniques.

Visual-based approaches.

Several works use visual features derived from the images' content and show that they are informative for predicting images' privacy settings^{7;12-14;24;27;33;89-102;191}. For example, Buschek et al.³³ presented an approach to assigning privacy to shared images using metadata (location, time, shot details) and visual features (faces, colors, edges). Zerr et al.⁷ proposed privacy-aware image classification and learned classifiers on Flickr images. The authors considered image tags and visual features such as color histograms, faces, edge-direction coherence, and Scale Invariant Feature Transform (SIFT) for the privacy classification task. Squicciarini et al.^{12, 13} performed an in-depth analysis of image privacy classification using Flickr images and found that SIFT and image tags work best for predicting privacy of users' images. Given the recent success of CNNs, several researchers^{14;24;27;89-91} showed promising privacy prediction results compared with visual features such as SIFT and GIST. Yu et al.⁴¹ adopted CNNs to achieve semantic image segmentation and also learned object-privacy relatedness to identify privacy-sensitive objects. Using CNNs, some works started to explore personalized privacy prediction models^{3;32;40}. In this context, it is worth mentioning that CNNs were also used in another body of privacy related work such as multi-party privacy conflict detection¹⁰⁵ and automatic redaction of sensitive image content⁹⁷.

Tag-based approaches.

Previous work in the context of tag-based access control policies and privacy prediction for images^{7;54;56;62;67;68;82-88} showed initial success in tying user tags with access control rules.

For example, Squicciarini et al.^{62, 88}, Zerr et al.⁷, and Vyas et al.⁸⁴ explored learning models for image privacy prediction using user tags and found that user tags are informative for predicting images’ privacy. Moreover, Squicciarini et al.⁶⁷ proposed an Adaptive Privacy Policy Prediction framework to help users control access for their shared images. The authors investigated social context, image content, and metadata as potential indicators of privacy preferences. Klemperer et al.⁶⁸ studied whether the user annotated tags help to create and maintain access-control policies more intuitively. However, the scarcity of tags for many online images²³ and the workload associated with user-defined tags preclude an accurate analysis of images’ sensitivity based on this dimension. Recently, in our prior work^{24;26;27}, we showed that the images’ tags that are automatically obtained from the visual content of images using Convolutional Neural Networks (CNNs) can improve the performance of image privacy prediction. Yet, since the CNNs are trained on ImageNet (1.2M+ images labeled with 1000 object categories)¹⁶ and Places2 (which contains 365 scene classes with 2.5 million images)²⁵, these tags depict objects or scenes given in the image and fail to capture the privacy characteristics (or orientation) of the image while generating the tags.

To this end, drawing ideas from collaborative filtering, we recommend privacy-aware tags for online images that have the potential to improve the set of user tags for online image sharing.

4.3 Privacy-Aware Image Tag Recommendation

Our approach to recommending privacy-aware tags for newly posted images in online content sharing sites is inspired from collaborating filtering (CF)¹⁹². Particularly, in user-item CF, items are recommended to users by finding the most similar users to the target user (from the user-item matrix) and recommending items to the target user based on the items that the similar users purchased/seen. The large amounts of images posted on the Web in recent years facilitate the study of potential relationships between images and tags. Our approach leverages these ideas to exchange tags between similar images. The analogy with conventional CF methods is that images correspond to users and tags correspond to items (i.e., in our

Algorithm 1 Tag Recommendation

```
1: Input: A dataset  $\mathcal{D} = \{I_1, \dots, I_n\}$  of images and their sets of tags  $\{T_1, \dots, T_n\}$ ; a target image  $I$  and its set of tags  $T$ ;  $pr(I)$  the privacy label of the target image  $I$  (could be private or public);  $k$  the number of nearest neighbors of  $I$  from  $\mathcal{D}$ ;  $r$  the number of tags to be recommended.
2: Output: A set  $R$  of recommended tags for  $I$ .
3:  $R \leftarrow \phi$ ; // the set of recommended tags, initially empty.
4:  $S \leftarrow \phi$ ;
5: if  $T = \phi$  then // if the set of tags is empty.
6:    $\mathbf{x} \leftarrow \text{ImageContentEncoding}(I)$ ; // deep features for  $I$ 
7:   for all  $I_j \in \mathcal{D}$  do
8:      $\mathbf{x}_j \leftarrow \text{ImageContentEncoding}(I_j)$ ; // deep features for  $I_j$ 
9:      $s_j \leftarrow \text{similarity}(\mathbf{x}, \mathbf{x}_j)$ ; // compute the visual content similarity between  $I$  and  $I_j$ 
10:     $S \leftarrow S \cup (I_j, s_j)$ ; // store  $I_j$  and its similarity with  $I$ 
11:   end for
12: else
13:    $\mathbf{x} \leftarrow \text{ImageTagEncoding}(I)$ ; // get tags' features of  $I$ 
14:   for all  $I_j \in \mathcal{D}$  do
15:      $\mathbf{x}_j \leftarrow \text{ImageTagEncoding}(I_j)$ ; // get tags' features of  $I_j$ 
16:      $s_j \leftarrow \text{similarity}(\mathbf{x}, \mathbf{x}_j)$ ; // compute the tags similarity between  $I$  and  $I_j$ 
17:      $S \leftarrow S \cup (I_j, s_j)$ ; // store  $I_j$  and its similarity with  $I$ 
18:   end for
19: end if
20:  $S.\text{similarities.sort}()$ ; // sort the images in decreasing order of their similarity scores
21:  $S \leftarrow$  top  $k$   $(I_j, s_j)$  entries; // get  $k$  images with the highest similarities with  $I$ , and their similarities
22:  $W \leftarrow \text{TagRanking}(S, pr(I))$ ; // rank the tags from  $S$  images
23:  $R \leftarrow r$  tags with the highest scores from  $W$ ;
24: return  $R$ 
```

setting, we deal with an image-tag matrix). Specifically, we aim to recommend tags for a target image by transferring privacy-aware tags from its most similar images, which are obtained from a large collection. We base our models on the assumption that *privacy-aware similar images possess similar tags*.

Algorithm 1 describes the process in detail. Specifically, the nearest neighbors of a target image are found by comparing rows in the image-tag matrix. Recommendations are made for the target image based on the neighboring images' tags (as a privacy-aware weighted sum of occurrences of tags). A common problem in CF is the *cold start* problem¹²⁴. In our case, this refers to images that have very few tags or no tags at all, and hence, there

is not enough information available to find accurate nearest neighbors for a target image. However, in our domain, images can be represented using two different views or feature types: (1) image content; and (2) image tags. We take advantage of both of these views (as shown in Algorithm 1).

The input of the algorithm is a dataset $\mathcal{D} = \{I_1, \dots, I_n\}$ of images and their associated sets of tags, $\{T_1, \dots, T_n\}$, respectively; a target image I and its set of tags T , which could possibly be empty; $pr(I)$ the privacy label of I , which could be *private* or *public*; k the number of nearest neighbors of I from \mathcal{D} ; and r the number of tags to be recommended. The output of the algorithm is a ranked list of r tags, which are recommended for the target image. The algorithm starts by checking if the set of tags T corresponding to the target image I is empty (Alg. 1, line 5). If $T \neq \phi$, the similarities between I and all images in $\mathcal{D} \setminus \{I\}$ are computed based on images’ tags (Alg. 1, lines 13-18). The top k most similar images to I are returned (Alg. 1, lines 20-21) and the candidate set that represents the union of the sets of tags extracted from these k similar images is ranked inside the subroutine for tag ranking (line 22). The tag ranking subroutine is described in Algorithm 2. The most highly ranked r tags from the candidate set are returned as recommended tags for the target image I (Alg. 1, line 23-24). For the cold start setting, if the initial tag set is empty, i.e., $T = \phi$ for image I , Algorithm 1 recommends r tags from the k most similar images in \mathcal{D} , where, this time, the similarity is computed based on image content features (not tags) (Alg. 1, lines 5-12).

For each tag in the candidate set, we compute its score as the privacy-aware sum of similarities between the target image and its similar images that contain that tag (Alg. 2, lines 6-12). This weighting method was employed based on the assumption that a “good” tag is very likely to be exchanged between similar images. Specifically, the weight (or score) of a tag t , w_t , is computed as:

$$w_t = \sum_{j \in \mathcal{S}} c_{jt} \cdot s_j \cdot P(t|pr(I)) \quad (4.1)$$

where \mathcal{S} represents the neighborhood of I , i.e., its k most similar images from \mathcal{D} , c_{jt} is an

Algorithm 2 Tag Ranking

```
1: function TagRanking( $S, pr(I)$ )
2:    $W \leftarrow \phi$ ; // the set of tags and their scores, initially empty.
3:   for all  $I_j \in S$  do
4:      $T_j \leftarrow I_j.tags$  // get the set of tags of image  $I_j$ .
5:      $s_j \leftarrow I_j.similarity$  // similarity of target image and  $I_j$ .
6:     for all  $t \in T_j$  do
7:        $w_t \leftarrow W.scoreOf(t)$  //  $w_t$  stores the score of  $t$ 
8:       if  $w_t = \text{null}$  then // if tag  $t$  is not in  $W$  already
9:          $W \leftarrow W \cup (t, 0)$  // add  $t$  to  $W$ 
10:      end if
11:       $w_t \leftarrow w_t + s_j \cdot P(t|pr(I))$  //score of  $t$  weighted by privacy
12:    end for
13:  end for
14:   $W.scores.sort()$  // sort the scores in  $W$  in the decreasing order.
15:  return  $W$ .
16: end function
```

indicator variable, which is 1 if tag t belongs to the tag set T_j of image I_j from \mathcal{S} and 0 otherwise, and s_j is the similarity between image I_j and I . The probability $P(t|pr(I))$ is the likelihood of the tag t belonging to one of the privacy classes (i.e., public or private) corresponding to the privacy of the target image I . For instance, if I is of private class, then $P(t|pr(I))$ gives the probability of tag t belonging to the set of private images. In experiments, the likelihood is calculated based on the dataset \mathcal{D} . We wish to obtain privacy-aware tags, i.e., tags weighted by their likelihood of occurrence in private or public classes, without missing out on the high-quality tags. Thus, we consider privacy-aware similarity that relies on the privacy likelihood of the tag instead of considering a privacy-enforced similarity. Here, we define privacy-enforced similarity as a similarity that considers privacy as an additional parameter in the image similarity, i.e., tags could be exchanged between images of the same privacy class (either public or private). A similarity weighted with privacy likelihood favors tags with a given privacy setting as opposed to the privacy-enforced similarity that would enforce tags of the same privacy settings as the target image. For example, using privacy-enforced similarity, for Figure 4.1(b) (given its public nature), tags such as “Women,” “Girl” (inclined to private class) would not be recommended. Conversely, privacy-aware weights can help obtain tags that are descriptive of an image content and help

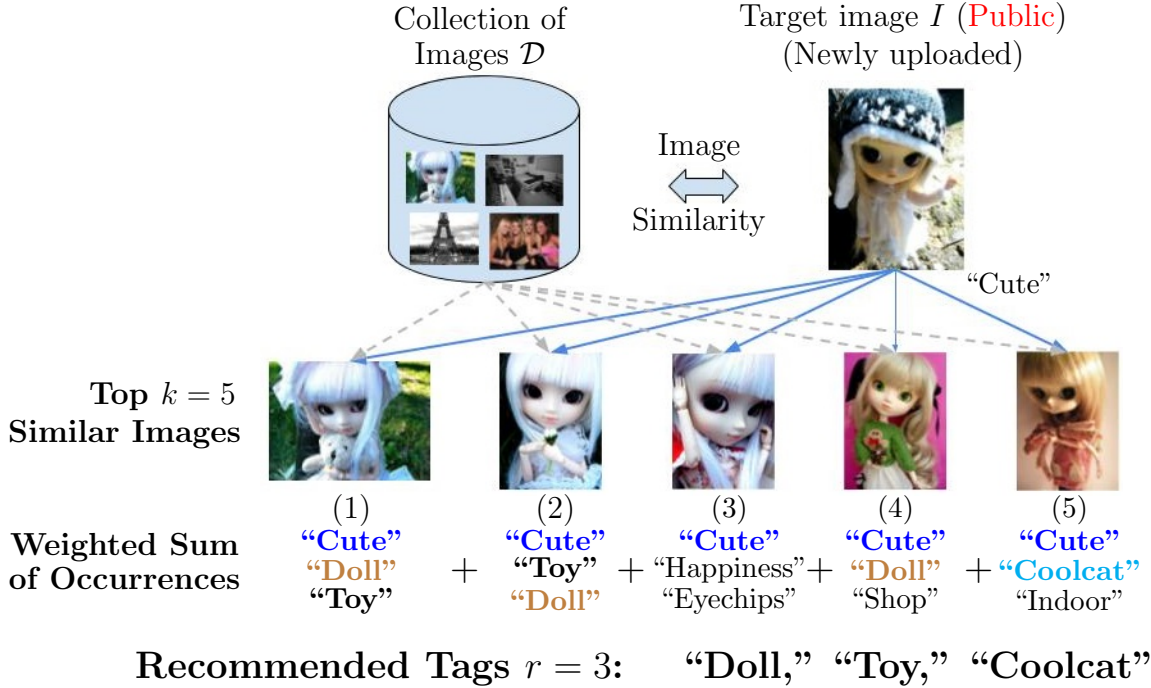


Figure 4.2: Illustration of the privacy-aware tag recommendation algorithm using an example: 1) A newly uploaded image on the Web that has an incomplete set of user-input tags, i.e., {“Cute”}, is considered as the target image I . 2) We can use images’ tags or content features to compute the similarity between the target image I and the images from the collection \mathcal{D} . For this example, we use visual content features to compute the similarity. 3) Top $r = 3$ tags { “Doll,” “Toy,” “Coolcat”} are recommended using top $k = 5$ similar images, through our privacy-aware tag recommendation approach. Note that the recommended tags “Doll” and “Toy” are appropriate tags for the target image I and can help correctly characterize its privacy class as public.

in identifying appropriate sharing needs of the image as it considers both image’s content and the privacy aspect of the image.

Figure 4.2 shows the illustration of the privacy-aware tag recommendation algorithm through an example. We consider a newly uploaded target image I on the Web that is of public class and has an incomplete set of user-input tags. For this illustration, we use visual content features to compute the similarity between the target image I and the images from the collection \mathcal{D} (shown in Figure 4.2 with a blue cylinder). Note that the images’ tags can be used to compute the similarity as well (as discussed in Alg. 1). The top $k = 5$ similar images are shown in the figure where the similarity decreases from left to right (the most similar image is labeled as (1)). Using these similar images, we obtain the set of candidate

Candidate Tags	Count	$P(t pr = private)$	$P(t pr = public)$	$w_t, s_j = 1$
Doll	3	0.1	0.9	$3 \times 0.9 = 2.7$
Toy	2	0.15	0.85	$2 \times 0.85 = 1.7$
Cute	5	0.7	0.3	$5 \times 0.3 = 1.5$
Coolcat	1	0.0	1.0	$1 \times 1.0 = 1.0$
Shop	1	0.0	1.0	$1 \times 1.0 = 1.0$
Eyechips	1	0.3	0.7	$1 \times 0.7 = 0.7$
Indoor	1	0.6	0.4	$1 \times 0.4 = 0.4$
Happiness	1	0.6	0.4	$1 \times 0.4 = 0.4$

Table 4.1: *Privacy-aware weighted sum of tag occurrences ($k = 5$) given that the target image is public. Bold words indicate the top $r = 3$ tags. Since the tag “Cute” appears already in the original set of user tags, we add the next important tag from the ranked list, i.e., “Coolcat.” The tags with same weights are selected randomly.*

tags for which we compute privacy-aware weights. The candidate tags and their privacy-aware weight calculation is shown in Table 4.1. For illustration purposes, we use $s_j = 1$ in Eq. 1, instead of the actual similarity between the target image and images in \mathcal{D} (where $0 \leq s_j \leq 1$). As we can see from Table 4.1, the tag “Cute” occurred in all five similar images, once in each image (see the column labeled as “Count”). The tag “Cute” is highlighted in blue color in Figure 4.2. Given that the target image is annotated as public, the tag “Cute” is weighted by the privacy likelihood $P(Cute|public)$, which is 0.3 (see Table 4.1). Recall that $P(t|private)$ and $P(t|public)$ are calculated from \mathcal{D} . Thus, the privacy-aware weighted sum of occurrences is given as 1.5. Table 4.1 shows the calculations for privacy-aware weighted sum of tag occurrences. Likewise, final weights are calculated for all candidate tags and top $r = 3$ tags are recommended for the target image (the recommended tags are shown in bold font in Figure 4.2 and Table 4.1). Note that since we consider privacy-likelihood of the tag instead of privacy-enforced similarity with the target image, the tag “Cute” describing the image content is recommended to the target image even though the tag “Cute” has privacy-related (“private”) connotations. However, since the tag “Cute” appears already in the original set of user tags, we do not add it to our set of recommended tags (to avoid over-counting), and add the next tag from the ranked list. We select the next tag with highest weight, i.e., “Coolcat” (shown in bold font in Table 4.1). The tags with the same

Dataset	#Total Images	#Avg. Tags	#min. Tags	#max. Tags	#Private Images	#Public Images
\mathcal{D}	8000	9.73	1	71	2000	6000
\mathcal{I}^E (Images I for Evaluation)	4189	18.65	11	78	1047	3142

Table 4.2: *Datasets summary.*

weights are selected randomly.

4.4 Dataset

We explore the effectiveness of the privacy-aware recommended tags for: (1) their ability to predict the private or sensitive content of online images; and (2) their relevancy to the images’ content. Hence, we evaluate our recommendation algorithm on Flickr images sampled from the PicAlert dataset, made available by Zerr et al.⁷. The PicAlert dataset contains both user-input tags and privacy labels. PicAlert is comprised of Flickr images on various subjects, which are manually labeled as *private* or *public*. The dataset contains photos uploaded in Flickr during the period from January to April 2010. The images have been labeled by six teams providing a total of 81 users of ages between 10 and 59 years. The guideline to select the label is given as: private images belong to the private sphere (like self-portraits, family, friends, someone’s home) or contain information that one would not share with everyone else (such as private documents). The remaining images are labeled as public. Each image was shown to at least two different users. In the event of disagreement, the photos were presented to additional users.

We split the PicAlert dataset into two subsets. The first subset corresponds to the dataset \mathcal{D} from Alg. 1 and is a collection of 8,000 images, labeled as private or public, that are used to recommend tags for the target images. We refer to this subset as \mathcal{D} . The second subset corresponds to target images that we use for evaluation and consists of 4,189 images from PicAlert, also labeled as private or public. We refer to this subset as \mathcal{I}^E or images I for evaluation. The ratio of public to private images in both the subsets \mathcal{D} and \mathcal{I}^E is 3 : 1. Table 4.2 shows a summary (number of total images, the average number of

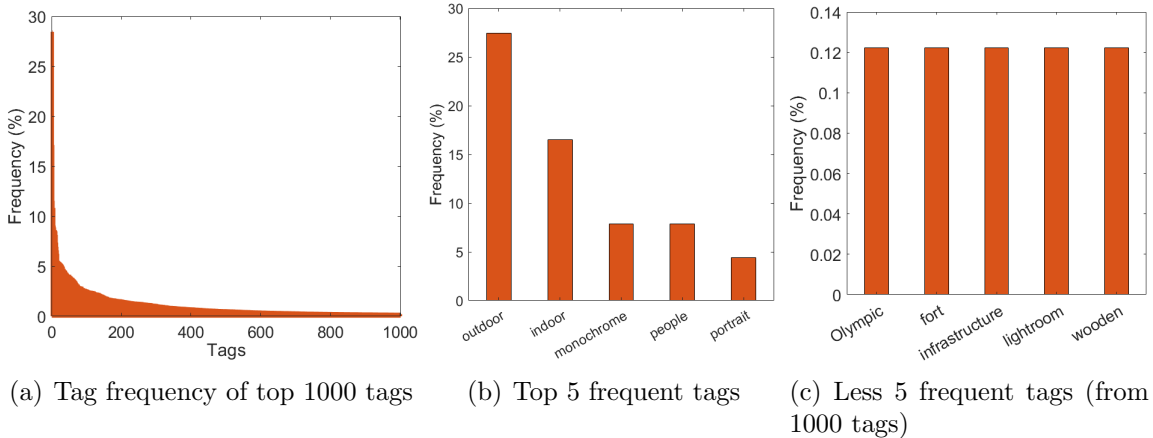


Figure 4.3: Tag frequency (%) in the PicAlert dataset. The frequencies are normalized by the size of the dataset.

tags per image, the minimum number of tags per image, the maximum number of tags per image, number of private and public images) of these datasets. For each image I in \mathcal{I}^E , we randomly split its set of tags into two subsets (i.e., *visible* and *hidden*). The motivation behind using random split is that newly uploaded images might have an incomplete and/or noisy set of user-input tags²³ and we desire to know if the proposed algorithm can overcome these challenges. The *visible* subset is denoted by T in Alg. 1 and is used to compute the similarity between the *visible* subset of the target image I in \mathcal{I}^E with the original set of tags of images in \mathcal{D} . The *hidden* subset is considered as gold standard for the evaluation of recommended tags. To calculate the precise similarity between two images using tags, we want to have at least five tags in the set of visible tags. Hence, we consider images with a number of user tags greater than 10 for the dataset \mathcal{I}^E (see Table 4.2, #minimum tags). In case less than 10 tags are available for an image, we can use the image content similarity. We filter out stop words, numbers, URLs, words with length less than 3 characters, and words with document frequency less than 2. After preprocessing, the size of the vocabulary is reduced to $\approx 19,000$. Note that for similarity computation (cosine in our experiments), we used the stemmed version of tags and synonyms obtained from WordNet¹⁹³. We also plot the frequency of top 1,000 tags normalized by the dataset size in Figure 4.3(a). The plot shows that top 200 tags befall in 3%–30% of the data with very few tags occurring in around

20% of the dataset. Note that most of the tags occur below 3% showing the variation in the images’ subjects and the complexity of our dataset. We also show the top 5 most frequent and less frequent tags with their frequency in Figures 4.3(b) and 4.3(c), respectively. Note that the frequencies of tags are normalized by the dataset size.

4.5 Experiments and Results

In this section, we evaluate the tags obtained by the privacy-aware tag recommendation algorithm for images in \mathcal{I}^E , by transferring tags from their most similar images from \mathcal{D} in several settings. That is, the quality of recommended tags is determined by: (1) *whether these tags hint to specific image privacy settings*; and (2) *whether these tags are good enough to describe the content of an image*. Hence, we adopt two evaluation mechanisms: (1) we examine the performance of models trained on the recommended tags combined with the original tags (when available) for image privacy prediction to determine their ability in building more robust models for identifying private or sensitive content for online image sharing; and (2) we compare the recommended tags against the ground-truth, i.e., the *hidden* set of tags, and also evaluate their quality through crowd-sourcing. We provide details of these evaluation types below.

Image Privacy Prediction. Similar to prior works on privacy prediction^{7;12;13;24;27}, we aim at identifying generic privacy patterns using the recommended tags to verify if these tags are indicative of the privacy classes. For this, we split \mathcal{I}^E into two subsets *Train* and *Test* to determine if the recommended tags are able to enhance the training set and learn better privacy characteristics. From \mathcal{I}^E , we randomly sample 3,689 images for *Train* and 500 for *Test*. We use *Train* to train Support Vector Machine (SVM) classifiers based on the recommended tags and use *Test* to test these classifiers. We provide the privacy class of images in *Train* as input to Alg. 1 and generate privacy-aware recommended tags for these images by exchanging tags from similar images in \mathcal{D} . The similarity between images is computed between the *visible* set of a target image in *Train* and all available tags from an image in \mathcal{D} . We train SVM classifiers on these recommended tags of *Train* and evaluate

them on the visible tags of the images in *Test*. Note that we do not recommend tags to images in *Test* as we assume that we do not know the privacy class of these images. We use the Weka implementation of SVM classifiers and choose the hyper-parameters that give best performance on *Train* using 10-fold cross-validation. For hyper-parameters, we experimented with $C \in \{0.001, 0.01, 1.0, \dots, 10\}$, kernels: Polynomial and RBF, the γ parameter in RBF, and the degree d of a polynomial.

Tag relevance. To evaluate the relevancy of the recommended tags, we randomly sample 500 images from \mathcal{I}^E . We denote this subset as *DRel*. We recommend privacy-aware tags for images in *DRel* by exchanging tags from similar images in \mathcal{D} . The similarity between images is computed between the *visible* set of a target image in *DRel* and all available tags from an image in \mathcal{D} . The *hidden* subset is considered as gold standard for evaluation, and contrasted with the predicted tag set. We also conduct a crowd-sourcing experiment to determine whether the recommended tags of the *DRel* dataset are relevant to the image’s content.

For all the experiments, we generate five random splits of visible and hidden subsets of tags and report performance (Accuracy, F1-measure, Precision, Recall) averaged over these five splits. We use a Boolean representation of tags, i.e., 1 if a tag is present for an image and 0 otherwise, since tags generally appear only once per image.

4.5.1 Evaluation of Privacy-Aware Recommended Tags by Privacy Prediction

The performance of privacy-aware recommended tags for image privacy prediction. We first evaluate our privacy-aware recommended tags obtained by the proposed weighting scheme in an ablation experiment for image privacy prediction. Specifically, we compare the performance of SVM classifiers trained only on recommended tags, where the recommended tags are obtained in several settings: (1) by our privacy-aware scoring mechanism, denoted as **p-Weights**, that ranks candidate tags using a privacy-aware weighted sum of tag occurrences (see Eq. 1); (2) recommending privacy-aware tags from the candidate

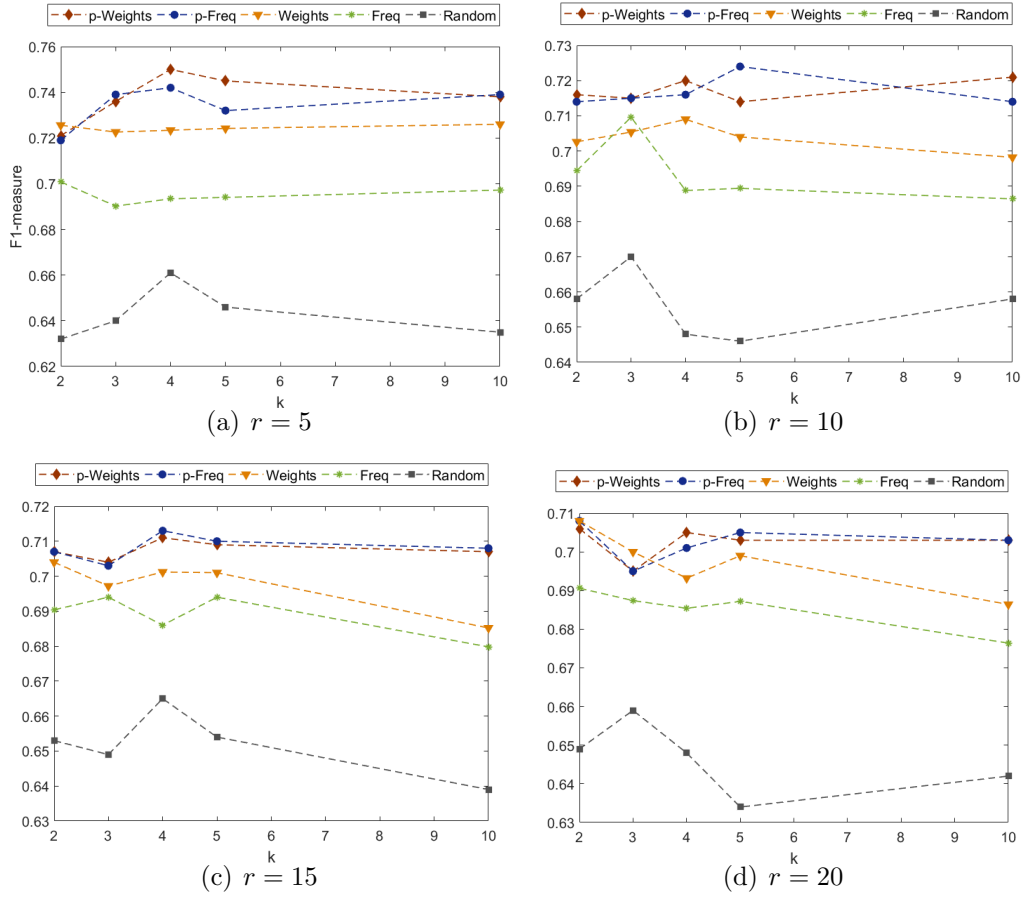


Figure 4.4: *F1-measure obtained for various parameter values, k and r of Alg. 1. p -Weights and p -Freq are privacy-aware scoring mechanism whereas Weights, Freq and Random are privacy-oblivious scoring mechanisms.*

set of tags based on their frequency in the k similar images, without considering images' similarity s_j (see Eq. 1 with $s_j = 1$), denoted as **p-Freq** (privacy-aware); (3) recommending tags by weighted sum of occurrences without considering the privacy likelihood (i.e., Eq. 1 without the term $P(t|pr(I))$), denoted as **Weights** (not privacy-aware); and (4) recommending tags based on their frequency in the k similar images, without considering images' similarity s_j and the privacy likelihood $P(t|pr(I))$, denoted as **Freq** (not privacy-aware). We also compare **p-Weights** with a random approach that recommends r tags randomly from the vocabulary of tags, denoted as **Random** (not privacy-aware).

To compare these methods, we study Algorithm 1 in the setting where each image in *Train* has a seed set of tags associated with it, i.e., $T \neq \phi$ (lines 13-18). The similarity between images is thus computed between the *visible* set of a target image in *Train* and all available tags from an image in \mathcal{D} . The similarity between two sets of tags is given as the cosine similarity of the corresponding bag-of-words vectors. We experiment with various numbers of similar images $k = 2, \dots, 10$, in steps of 1, and recommended tags $r = 5, \dots, 20$, in steps of 5.

Figure 4.4 shows the average F1-measure achieved by SVM classifiers using the four ranking strategies for different values of k (number of similar images) and r (number of recommended tags), and the Random naive approach. The SVMs are trained on the recommended tags of the *Train* dataset and evaluated on the visible tags of the *Test* dataset. We can see from the figure that recommended tags obtained using **p-Weights** can learn better privacy characteristics than **Random**, **Weights**, **Freq** (not privacy-aware) and perform comparable to **p-Freq** (privacy-aware), for values of $r = \{10, 15, 20\}$ regardless of the value of k . We also notice that the **p-Weights** scoring mechanism achieves the best performance for $r = 5$ and $k = 4$, outperforming all the other models including **p-Freq**, which shows that all scoring components (s_j and $P(t|pr(I))$) play a role in the overall performance. It is also interesting to mention that **Weights** (not privacy-aware) scoring mechanism consistently performs better than **Freq** (not privacy-aware) scoring method.

In the previous experiment, we only used recommended tags to compare various scoring schemes. Next, we wish to identify how recommended tags perform when we add them to

the visible set of images in *Train* for privacy prediction. In what follows, since the results are generally similar for $k = 4, 5, 10$ (see Figure 4.4), we use these values to augment the set of tags for *Train* with recommended tags by **p-Weights**.

Features	Acc. %	F1	Precision	Recall
<i>vt</i>	74.83	0.743	0.739	0.748
$k = 4$				
<i>vt</i> & <i>rt</i> ($r = 5$)	77.84	0.766	0.755	0.778
<i>vt</i> & <i>rt</i> ($r = 10$)	77.47	0.763	0.752	0.776
<i>vt</i> & <i>rt</i> ($r = 15$)	77.31	0.757	0.744	0.771
<i>vt</i> & <i>rt</i> ($r = 20$)	76.83	0.754	0.741	0.769
$k = 5$				
<i>vt</i> & <i>rt</i> ($r = 5$)	77.96	0.769	0.758	0.781
<i>vt</i> & <i>rt</i> ($r = 10$)	77.80	0.766	0.755	0.778
<i>vt</i> & <i>rt</i> ($r = 15$)	77.60	0.764	0.752	0.776
<i>vt</i> & <i>rt</i> ($r = 20$)	77.27	0.760	0.747	0.773
$k = 10$				
<i>vt</i> & <i>rt</i> ($r = 5$)	78.20	0.772	0.762	0.783
<i>vt</i> & <i>rt</i> ($r = 10$)	77.80	0.765	0.754	0.777
<i>vt</i> & <i>rt</i> ($r = 15$)	77.92	0.767	0.758	0.778
<i>vt</i> & <i>rt</i> ($r = 20$)	77.43	0.758	0.745	0.771

Table 4.3: Performance for privacy prediction after adding recommended tags. “*vt*” denotes a set of visible tags and “*rt*” denotes a set of recommended tags, e.g., {“cute”, “toy”, “doll”}. “ r ” is the number of tags recommended.

The performance of privacy-aware recommended tags for image privacy prediction when added to the visible tags. Table 4.3 shows the performance (Accuracy, F1-measure, Precision, Recall) obtained by the SVM classifiers trained on the combination of recommended tags (*rt*) and visible tags (*vt*) (as we increase r from 5 to 20) for the images in *Train* and evaluated on the fixed set of visible tags of the images in *Test* (for consistency). The results show that the performance of privacy prediction improves when we add recommended tags to the set of visible tags for images in *Train*. Specifically, we get the best performance when we use $k = 10$ and $r = 5$ with F1-measure of 0.772, whereas the SVM trained on only visible tags achieves 0.743 F1-measure, yielding an improvement of 3% in overall performance. We notice that generally, the performance increases with the decreasing value of r (best performance is given by $r = 5$) and increasing value of k (best

performance is given by $k = 10$). This can be justified by the fact that given the diverse nature of the data and the large vocabulary, a large r may introduce noise in the results. Similarly, a high value of k leads to higher number of similar images from which we get a set of good candidate tags.

The previous experiments used image tags to find the neighborhood of an image. However, not all images on social networking sites have user tags associated with them²³, and this gives rise to the cold start problem for collaborative filtering. Next, we discuss how we overcome the problem. In the following experiments, we use the privacy-aware weighting scheme **p-Weights**, $k = 10$ and $r = 5$.

4.5.2 Solution to the Cold Start Problem

Cold start is a challenging problem particularly in many collaborative filtering approaches, where the absence of items (i.e., tags, in our case) that are used to bootstrap the algorithms may theoretically hinder the recommendations to be produced. Hence, we evaluate our approach **p-Weights** for image tag recommendation in the setting where we assume that each image in *Train* has no tags associated with it, i.e., $T = \phi$. This involves recommending tags from visually similar images (lines 5-12 of Alg. 1). The similarity between two images is given as the cosine similarity of the corresponding feature vectors. We consider two types of image features extracted from a deep convolutional neural network (CNN): 1) *deep visual feature pool₅*, and 2) *deep tags*. The choice of the features is motivated by their performance for privacy prediction in prior works^{14;24;27}.

We extract the deep visual features and deep image tags using GoogLeNet architecture¹⁸, which implements a 22 layer deep network with the Inception architecture. The architecture is a combination of all layers with their output filter bank concatenated to form input for the next stage. We extract visual features *pool₅* from the layer named as “pool₅/drop_7x7_s1” (dropout layer). For deep tags, we use the probability distribution over 1,000 object categories for the input image obtained by applying the softmax function over the last fully-connected layer of the CNN. We consider the top k objects of highest probabilities as *deep*

tags. The GoogLeNet network is pre-trained on a subset of the ImageNet dataset¹⁶, which is distributed with the CAFFE open-source framework for CNN¹⁰⁷.

Features	Acc.%	F1	Precision	Recall
<i>rt</i> -pool ₅	75.74	0.743	0.729	0.757
<i>rt</i> -DT	74.19	0.731	0.725	0.742
<i>vt</i>	74.83	0.743	0.739	0.748
DT	68.54	0.645	0.619	0.685

Table 4.4: Visual content-based similarity ($k = 10$, $r = 5$).

Table 4.4 shows the performance of privacy prediction obtained by the SVM models trained on the privacy-aware tags recommended from visually similar images based on pool₅ (*rt*-pool₅) and deep tags (*rt*-DT) for the images in *Train* and evaluated on the visible tags of the images in *Test*. For this experiment, we assume that we do not know the set of visible tags for images in *Train*. However, we wish to examine how would the recommended tags obtained using visual content similarity perform as compared to the visible tags and predicted deep tags (DT) of images in *Train*, as done in our prior work^{24;27}. Thus, we also show the performance of the models trained on visible tags alone (*vt*) and deep tags (DT) in Table 4.4. The results show that the models trained on the recommended tags yield similar results to the models trained on visible tags (user-input tags - if we would know them) and outperform those trained on the top k predicted deep tags (from GoogLeNet) for each image in *Train*^{24;27}. Precisely, we obtain maximum value of F1-score as 0.743 and best recall of 0.757 with recommended tags $r = 5$.

From the table, we observe that the models trained on tags recommended from visually similar images calculated based on pool₅ (*rt*-pool₅) outperform those trained on tags recommended from visually similar images calculated based on deep tags (*rt*-DT). The models trained on recommended tags obtained using pool₅ also outperform the models trained on the top k predicted deep tags (DT) presented in our prior works^{24;27}, that are generated without any tag recommendation (i.e., the exchange of tags from similar images). This can be explained by the fact that the deep tags belong to only 1,000 object categories due to which many relevant tags can not be captured. For example, tags such as “walking” and

“culture” are not present in the 1,000 object categories, but may be relevant tags for a given picture.

4.5.3 The Proposed Approach vs. Prior Privacy Prediction Works

We compare the performance of privacy prediction models trained on the user tags improved by the set of recommended tags with the performance obtained by following prior privacy prediction approaches. Mainly, we compare the performance obtained with the recommended tags with two types of features, viz., visual features (fc_8 and PCNH) and tag features (User Tags, Deep Tags, and their combination).

1. fc_8 ^{24;27}: We consider the model trained on the features extracted from the last fully-connected layer of AlexNet, i.e., fc_8 as our baseline, since in our previous work we achieved a good performance using these features for privacy prediction.

2. **PCNH privacy framework**¹⁴: This framework combines features obtained from two architectures: one that extracts convolutional features (size = 24, referred as Convolutional CNN), and another that extracts object features (size = 24, referred as Object CNN). The Convolutional CNN contains two convolutional layers and three fully-connected layers of size 512, 512, 24, respectively. On the other hand, the object CNN is an extension of AlexNet architecture that appends three fully-connected layers of size 512, 512, and 24, at the end of the last fully-connected layer of AlexNet and forms a deep network of 11 layers. The two CNNs are connected at the output layer. The PCNH framework is first trained on the ImageNet dataset and then fine-tuned on a small privacy dataset.

3. **Image Tags**: Previous works used user tags (UT)^{7;12}, deep tags (DT)^{24;27} and their combination (UT+DT)^{24;27} for privacy prediction and hence, we consider models trained on these tags as other baselines. Note that we describe deep tags in details in our previous experiment where we evaluated the cold start problem.

Table 4.5 compares the privacy prediction performance obtained with the recommended tags (RT) with the performance obtained by the prior works. The table shows that when we add the recommended tags (RT) to the existing user tags (UT), the F1-measure improves

Features	Acc.%	F1	Precision	Recall
With Recommended Tags				
RT	75.37	0.75	0.747	0.754
UT+RT	78.20	0.772	0.762	0.783
UT+DT+RT	81.9	0.810	0.811	0.819
Visual features				
fc ₈ ^{24;27}	81.16	0.805	0.803	0.812
PCNH ¹⁴	77.91	0.768	0.764	0.779
Tag features				
UT ^{7;12;13}	74.83	0.743	0.739	0.748
DT ^{24;27}	68.54	0.645	0.619	0.685
UT+DT ^{24;27}	78.81	0.786	0.784	0.789

Table 4.5: Comparison of privacy prediction performance obtained using the proposed approach and prior privacy prediction approaches.

by 3% over the user tags alone. Similarly, when we add the recommended tags (RT) to the combination of user tags and deep tags (UT + DT), we get improvement in the F1-measure of 3% over the combination of user tags and deep tags. We also observe that the model trained on the tag features with the recommended tags (UT+DT+RT) yields a better performance to the models trained on the visual features fc₈ and PCNH. For example, the UT+DT+RT achieves an F1-measure of 0.81, whereas fc₈ and PCNH obtain F1-measure of 0.805 and 0.768, respectively. Even though the tag features do not yield a great improvement over visual features (fc₈), tag features are also essential for the privacy prediction as they provide other aspects of an input image that have not been captured by the visual content. For example, consider an image containing “people with glasses in their hands.” Solely using visual content, one cannot differentiate from a “birthday party” to “event launch party.” User tags (generated by image owner) can contain such information, which can provide relevant cues for privacy prediction. It is interesting to mention here that, improving user tags with the set of recommended tags reduces the performance gap between the tag and visual features. Visual features and tag features can complement each other, and hence, can be combined to obtain improved privacy prediction performance in the future. Additionally, these privacy-aware tags can predict privacy of an image accurately even when access to the visual content of the image is not allowed due to users reluctance to share the actual image

for visual content analysis (which could reveal a user’s identity through the face and friends, etc.).

Next, we compare the privacy prediction performance of recommended tags by our approach (privacy-aware) with the tags generated by prior image annotation mechanism (not privacy-aware).

4.5.4 The Proposed Approach vs. Prior Image Annotation Works

In this experiment, we compare the performance of privacy prediction using tags recommended by the proposed approach “p-Weights” against the tags recommended by prior nearest neighbors based image annotation works. Particularly, we consider the modified version of “Fast image tagging” (or FastTag)¹²² and image annotation by Makadia et al.¹²¹ as our baselines. We provide details of our baselines as follows.

1. FastTag¹²²: FastTag addresses the tag sparsity problem, which motivates the choice of FastTag as our baseline. This is particularly critical to our dataset as we can see in Figure 4.3, very few tags occur in around 20% of the dataset. Additionally, similar to our approach, FastTag also considers images with partial tags to predict tag annotations. For FastTag, authors considered traditional image features such as Gist descriptor³⁵, global color histograms, and bag-of-word visual features. Recently, Mayhew et al.¹⁹⁴ trained nearest neighbors-based image annotation algorithms using the features derived from CNNs and achieved better performance than using traditional image features. Thus, similar to our approach, we use pool₅ (CNN based feature representation) as image features in the FastTag algorithm. For other parameters in FastTag, we consider the best (default) values given by the authors.

2. Makadia et al.¹²¹: Similar to our work, Makadia et al.¹²¹ also transfers tags from the most similar images of a target image and thus, we consider it as our another baseline. The tag transfer mechanism of Makadia et al.¹²¹ is different from our tag scoring mechanism “p-Weights”. Makadia et al.¹²¹ follows a three step process to transfer tags to a target image from its neighbors. First, the authors rank the tags according to their frequency in the dataset

Features	Acc.%	F1	Precision	Recall
#1 Original User Tags (Visible Tags)				
vt	74.83	0.743	0.739	0.748
#2 Prior Image Annotation Works				
vt & rt by FastTag ¹²²	74.55	0.741	0.738	0.745
vt & rt by Makadia et al. ¹²¹	74.87	0.730	0.723	0.749
#3 Visual Content Similarity ($T = \phi$)				
vt & rt ($r = 5$)	75.23	0.741	0.730	0.752
vt & rt ($r = 10$)	75.63	0.742	0.727	0.757
vt & rt ($r = 15$)	76.71	0.752	0.737	0.768
vt & rt ($r = 20$)	76.27	0.747	0.732	0.763
#4 Tag Similarity ($T \neq \phi$)				
vt & rt ($r = 5$)	78.20	0.772	0.762	0.783

Table 4.6: *Privacy-aware Tag recommendation vs. Prior Image Annotation Works.*

(in our case \mathcal{D}). Second, the highest ranking tags of the first neighbor (first similar image) are transferred to the target image. If the number of tags of the first neighbor is greater than r , then only the top r tags are transferred. Last, the tags of neighbors $2, \dots, k$ are ranked based on two factors: 1. co-occurrence of tags in training (\mathcal{D}) with the tags transferred in step 2; and 2. frequency of tags of neighbors $2, \dots, k$. The highest ranking tags are selected and the remaining tags (r - tags transferred in step 2.) are transferred to the target image. Makadia et al.¹²¹ also considered color and texture-based visual features (traditional image features). Even in this case, we use pool_5 as image features for an unbiased comparison. Similar to our approach, we compute cosine similarity between two visual feature vectors to obtain top $k = 10$ neighbors and recommend $r = 5$ tags.

For this comparison, we include the tags obtained by both the settings when the seed set $T \neq \phi$ (tag similarity, Alg. 1, lines 13-18) and $T = \phi$ (visual content similarity, Alg. 1, lines 5-12). Specifically, we compare the models for privacy prediction trained on the combination of visible tags and recommended tags by the proposed approach with the models trained on the combination of visible tags and the tags obtained by FastTag.

Table 4.6 shows the privacy prediction performance comparison between the models trained on the combination of visible tags (vt) and the recommended tags (rt) by Alg. 1, FastTag and Makadia et al.¹²¹. From the table, we can observe that the models trained

on the tags obtained by the proposed approach perform better than the models trained on the tags obtained by FastTag and Makadia et al. ¹²¹. Specifically, models trained on the combination of visible tags and tags recommended by visual content yield F1-measure as high as 0.752 (Table 4.6, #3 Visual Content Similarity), whereas the models trained on the combination of visible tags and tags obtained by FastTag and Makadia et al. ¹²¹ get F1-measure of 0.741 and 0.730 respectively. We achieve the best performance of 0.772 (F1-measure) using the tag similarity (Table 4.6, #4 Tag Similarity). Note that the F1-measure obtained by models trained on the combination of visible tags and the tags obtained by FastTag (0.741) or Makadia et al. ¹²¹ (0.730) (Table 4.6, #2 Prior Image Annotation Works) is even slightly worse than the F1-measure (0.743) obtained for the models trained on only visible tags (Table 4.6, #1 Original User Tags). The results show that even though we use the same set of visual features (deep features) for all the three methods (p-Weights, FastTag and Makadia et al. ¹²¹) to generate the tags, the tags obtained by FastTag and Makadia et al. ¹²¹, which are privacy-oblivious, are not very helpful for identifying images’ private content. Despite that FastTag performs well for general image annotation ¹⁷⁹, it fails to recommend privacy preserving tags on the PicAlert dataset because, unlike our approach, the impact of the privacy of an image is not considered.

4.5.5 Quality Assessment of Recommended Tags

In the above experiments, we compared the privacy prediction performance obtained by privacy-aware and privacy-oblivious tags. In this experiment, we determine which set of recommended tags (privacy-aware vs. privacy-oblivious) describe an image’s content appropriately. Precisely, we obtain two sets of recommended tags: (1) using our privacy-aware weighting scheme, referred as privacy-aware tags (see Eq. 4.1), and (2) using weighting scheme without privacy likelihood, referred as privacy-oblivious tags (Eq. 4.1 without the term $P(t \text{—} \text{pr}(I))$). We compare these tags against the ground-truth (i.e., the *hidden* set of tags). For this experiment, we recommend tags (using both privacy-aware and privacy-oblivious weighting schemes) for images in *DRel*, where each image has a seed set of tags

	<u>User-input Tags</u>		<u>Recommended</u>
	<u>Visible</u>	<u>Hidden</u>	<u>Tags</u>
	Beauty	Geisha	People
	Light	Kyoto	Japan
	Travel	Japan	Asia
	Couple	Kimono	Geisha
	Woman	Traditional	Kimono
	Vintage	Asia	Kyoto
		People	Traditional
			<i>Culture</i>
		<i>street</i>	
		<i>walking</i>	

Figure 4.5: Image with recommended tags, $r=10$.

associated with it, i.e., $T \neq \phi$. For each image in $DRel$, we randomly split its set of tags into two subsets, i.e., visible and hidden, where the visible set is used for tag similarity and the hidden set is used as gold-standard set. The similarity between images is thus computed between the *visible* set of a target image in $DRel$ and all available tags from an image in \mathcal{D} .

Table 4.7 shows the performance (Precision@ r) obtained for $r \in \{1, \dots, 10\}$ tags recommended for the images in $DRel$ when compared against the gold-standard set of tags (those that are hidden from the original user tags). We compute Precision as the total number of *recommended* and *relevant* tags over the number of tags recommended (i.e., r). The results show that the privacy-aware tags obtain better precisions than the privacy-oblivious tags, yielding the highest precision of 0.197 ($r = 4$) using gold-standard. The gold-standard set is nothing but a subset of user annotated tags, which may not provide all the possible tags that can be associated with an image content. Hence, the gold-standard set may fail to capture highly relevant tags provided by the recommendation strategy. For example, in Figure 4.5, we can see that tags relevant to the image content (shown in italic) are recommended, but do not appear in the user-input tags. Specifically, even though tags such as *culture*, *street*, *walking* are consistent with the image content, these tags are not considered for calculating the precision values since they do not appear among the tags in the *hidden* set or gold-standard set.

Crowd-sourcing can be used to address the above limitation. Hence, we employ crowd-sourcing to make use of the “wisdom of the crowd,” as follows: we use two annotators from

#Tags (r)	Gold-standard		Crowd-sourcing	
	PA P@r	PO P@r	PA P@r	PO P@r
1	0.162	0.182	0.87	0.863
2	0.186	0.182	0.85	0.84
3	0.195	0.180	0.812	0.822
4	0.197	0.186	0.791	0.793
5	0.190	0.184	0.77	0.77
6	0.184	0.178	0.753	0.75
7	0.174	0.169	0.742	0.738
8	0.168	0.164	0.731	0.72
9	0.162	0.158	0.72	0.71
10	0.156	0.153	0.71	0.704

Table 4.7: Gold-standard and User evaluation of privacy-aware and privacy-oblivious recommended tags.

Figure Eight¹ to determine if the recommended tags are relevant to images’ content. For each tag, annotators were asked to choose between: *relevant*, *irrelevant* and *not sure*. To calculate precision values, we consider a tag as *Relevant* if at least one annotator marked it as *relevant* as the tags can be subjective and one annotator can observe more in an image than the other.

Table 4.7 also shows the performance obtained through crowd-sourcing. We notice that the results of crowd-sourcing are higher than those obtained by relying only on *gold standard* to compute the performance. Precisely, through crowd-sourcing, the precision increased from 0.197 (gold-standard set) to 0.87 for privacy-aware tags, reassuring that the generated tags are relevant to images’ content. Similarly, for privacy-oblivious tags, the precision increased from 0.182 to 0.863. The difference in the results can be justified by the fact that the user tags tend to be noisy, incomplete, and may not relate to the image content²³. We observe that, for the crowd-sourcing experiment, precision obtained using privacy-aware tags is higher than the precision obtained using privacy-oblivious tags for $r = \{1, 2, 7 - 10\}$. Note that for r ranging from 3 to 6, the performance of privacy-aware tags is comparable to the performance of privacy-oblivious tags. One reason could be that some relevant tags have higher weights and are recommended irrespective of their privacy likelihood. Consider

¹<https://make.figure-eight.com/>

#Tags	Nouns Only		Verb Only		Adjective Only		Noun & Verb	
	PA P@r	PO P@r	PA P@r	PO P@r	PA P@r	PO P@r	PA P@r	PO P@r
1	0.812	0.808	0.64	0.626	0.872	0.865	0.815	0.805
2	0.792	0.796	0.633	0.626	0.849	0.848	0.792	0.790
3	0.778	0.776	0.638	0.626	0.848	0.846	0.780	0.780
4	0.756	0.750	0.638	0.626	0.848	0.846	0.755	0.753
5	0.741	0.735	0.638	0.626	0.848	0.846	0.742	0.736
6	0.737	0.730	0.638	0.626	0.848	0.846	0.735	0.729
7	0.735	0.728	0.638	0.626	0.848	0.846	0.734	0.726
8	0.734	0.727	0.638	0.626	0.848	0.846	0.733	0.726
9	0.734	0.727	0.638	0.626	0.848	0.846	0.732	0.725
10	0.734	0.727	0.638	0.626	0.848	0.846	0.726	0.731

Table 4.8: User evaluation of recommended tags that are Noun, Verb, Adjective, and Noun & Verb. Privacy-aware tags are denoted as PA and privacy-oblivious are denoted as PO.

a private image of “people on the beach” for which “beach” (being considered as nature) would be recommended even though it has higher likelihood towards the public class.

The tags depicting objects (such as beach, furniture) or actions (such as walking) in images are more objectively identified by annotators, whereas abstract tags such as “beautiful,” “pretty,” etc., are more subjective. This could be another justification for the similar results that we obtain for privacy-aware and privacy-oblivious tags for values of $r = \{3-6\}$ in Table 4.7. To understand this, we further investigate both privacy-aware and privacy-oblivious sets of tags by obtaining part-of-speech (POS) tags for the recommended tags. The recommended tags for *DRel* contain approximately 45% of nouns, 4% of verbs, 5% of adjective POS tags, and the remaining are the proper nouns (44%). Table 4.8 shows the user evaluation of the recommended tags (privacy-aware and privacy-oblivious) that are nouns, verbs, adjective, and nouns & verbs. Note that we do not consider proper nouns as solely from the visual content (without user’s information), it is difficult to identify whether a particular place or a person is relevant to a target image. For example, one can recognize a “beach” from the visual content of a target image, but for some images it is difficult to know the exact location (i.e., a proper noun) of the beach (e.g., oregon coast). In the table, the privacy-aware tags are denoted as “PA”, and privacy-oblivious tags are denoted as “PO”. The table shows that for nouns (that depict objects and scenes in the image), privacy-aware tags obtain higher



Figure 4.6: *Subjective Adjective (Tags)*

performance than the privacy-oblivious tags for almost all values of r . Similarly, for verbs (that depict actions of the objects in the image), privacy-aware tags yield higher performance than the privacy-oblivious tags. Note that images might not have more than 1–2 verbs; thus the performance does not change after $r = 3$. Conversely, for adjectives, we observe that the performance is comparable for both sets of tags. One reason might be that the adjectives are subjective and even though the privacy-aware tags have recommended good adjective tags, those are not reflected in the performance for values of $r = \{3–6\}$ in Table 4.7. To illustrate this, we provide some examples in Figure 4.6 that contain subjective adjectives (tags). For example, for image (a), some people might identify that the shot was taken beautifully, and hence, they might consider tag “beautiful” as relevant tag for the image. On the other hand, others might find the animal scary and they might not consider the tag relevant.

4.6 Chapter Summary and Future Directions

We proposed an approach to recommending privacy-aware image tags that can improve the original set of user tags and, at the same time, preserve images’ privacy to help reduce the private content from the search results. Our approach draws ideas from collaborative filtering (CF). Although the user-input tags are prone to noise, we were able to integrate them in our approach and recommend accurate tags. More importantly, we simulated the recommendation strategy for newly-posted images, which had no tags attached. This is a

particularly challenging problem, as in many CF approaches, the absence of items (tags in our case) may theoretically hinder the recommendations to be produced, due to the lack of enough information available to find similar images to a target image. Through our experiments, we showed that we achieve better performance for image privacy prediction with recommended tags than the original set of user tags, which in turn indicates that the suggested tags comply to the images' privacy. We also show that improving user tags with a set of privacy-aware recommended tags can reduce the performance gap between the tag and visual features for privacy prediction. Visual features and tag features can complement each other, and hence, can be combined to obtain improved privacy prediction performance in the future. Last, we conducted a user evaluation to inspect the quality of our privacy-aware recommended tags. The results show that the proposed approach is able to recommend highly relevant tags.

In future work, it would be interesting to study the algorithm for multiple sharing needs of the user such as friends, family, and colleagues by considering privacy likelihood with respect to multi-class privacy settings. We plan to explore alternative ways of computing images' similarity, such as combining information from both tags and visual content. Also, another interesting direction would be to explore image-content features depicting various image subjects such as scene and location, which could lead to more accurate results.

Chapter 5

Dynamic Deep Multi-modal Fusion for Image Privacy Prediction

With millions of images that are shared online on social networking sites, effective methods for image privacy prediction are highly needed. In this chapter, we propose an approach for fusing object, scene context, and image tags modalities derived from convolutional neural networks for accurately predicting the privacy of images shared online. Specifically, our approach identifies the set of most competent modalities on the fly, according to each new target image whose privacy has to be predicted. The approach considers three stages to predict the privacy of a target image, wherein we first identify the neighborhood images that are visually similar and/or have similar sensitive content as the target image. Then, we estimate the competence of the modalities based on the neighborhood images. Finally, we fuse the decisions of the most competent modalities and predict the privacy label for the target image. Experimental results show that our approach predicts the sensitive (or private) content more accurately than the models trained on individual modalities (object, scene, and tags) and prior privacy prediction works. Additionally, our approach outperforms the state-of-the-art baselines that also yield combinations of modalities.

5.1 Introduction

Technology today offers innovative ways to share photos with people all around the world, making online photo sharing an incredibly popular activity for Internet users. These users document daily details about their whereabouts through images and also post pictures of their significant milestones and private events, e.g., family photos and cocktail parties¹. Furthermore, smartphones and other mobile devices facilitate the exchange of information in content sharing sites virtually at any time, in any place. Although current social networking sites allow users to change their privacy preferences, this is often a cumbersome task for the vast majority of users on the Web, who face difficulties in assigning and managing privacy settings². Even though users change their privacy settings to comply with their personal privacy preference, they often misjudge the private information in images, which fails to enforce their own privacy preferences³. Thus, new privacy concerns⁴ are on the rise and mostly emerge due to users' lack of understanding that semantically rich images may reveal sensitive information^{3;5-7}. For example, a seemingly harmless photo of a birthday party may unintentionally reveal sensitive information about a person's location, personal habits, and friends. Along these lines, Gross and Acquisti⁸ analyzed more than 4,000 Carnegie Mellon University students' Facebook profiles and outlined potential threats to privacy. The authors found that users often provide personal information generously on social networking sites, but they rarely change default privacy settings, which could jeopardize their privacy. Employers often perform background checks for their future employees using social networking sites and about 8% of companies have already fired employees due to their inappropriate social media content⁹. A study carried out by the Pew Research center reported that 11% of the users of social networking sites regret the content they posted¹⁰.

Motivated by the fact that increasingly online users' privacy is routinely compromised by using social and content sharing applications¹⁵, recently, researchers started to explore machine learning and deep learning models to automatically identify private or sensitive content in images^{3;7;12;14;24;26;27}. Starting from the premise that the objects and scene contexts present in images impact images' privacy, many of these studies used objects, scenes, and

Single modality is correct			
Image	Tags	Probabilities	
		Base classifiers	Fusion
	bed, studio dining table speakers, music	scene: 0.62 object: 0.5 tags: 0.29	Feature-level: 0.21 Decision-level: 0.33
(a)			
	birthday night party, life	scene: 0.57 object: 0.78 tags: 0.39	Feature-level: 0.21 Decision-level: 0.33
(b)			
	toEIC, native speaker, text document, pen	scene: 0.02 object: 0.15 tags: 0.86	Feature-level: 0.27 Decision-level: 0.33
(c)			
Multiple modalities are correct			
Image	Tags	Probabilities	
		Base classifiers	Fusion
	girl, baby indoor, people canon	scene: 0.49 object: 0.87 tags: 0.97	Feature-level: 0.77 Decision-level: 0.67
(d)			
	people, party awesome, tea bed, blanket	scene: 0.92 object: 0.38 tags: 0.7	Feature-level: 0.69 Decision-level: 0.67
(e)			
	indoor, fun party people	scene: 0.92 object: 0.73 tags: 0.77	Feature-level: 0.89 Decision-level: 1
(f)			

Figure 5.1: Anecdotal evidence of private images and their tags. The feature-level fusion is given as the concatenation of all the features (object, scene, tag) and the decision-level fusion is obtained by averaging the predictions.

user tags, or their combination (i.e., feature-level or decision-level fusion) to infer adequate privacy classification for online images.

However, we conjecture that simply combining objects, scenes and user tags modalities using feature-level fusion (i.e., concatenation of all object, scene and user tag features) or decision-level fusion (i.e., aggregation of decisions from classifiers trained on objects, scenes and tags) may not always help to identify the sensitive content of images. Figure 5.1 illustrates this phenomenon through several images. For example, let us consider image (a) in the figure. Both feature-level and decision-level fusion models yield very low private class

probabilities (Feature-level fusion: 0.21 and decision-level fusion: 0.33). Interestingly, a model based on the scene context (bedroom) of the image outputs a high probability of 0.62, showing that the scene based model is competent to capture the sensitive content of the image on its own. Similarly, for the image (b) (self-portrait) in Figure 5.1, where scene context is seldom in the visual content, the objects in the image (the “persons,” “cocktail dress”) are more useful (0.78) to predict appropriate image’s privacy. Moreover, for images such as “personal documents” (image (c)), user-annotated tags provide broader context (such as type and purpose of the document), capturing the sensitive content (0.86), that objects and scene obtained through images’ content failed to capture. On the other hand, in some cases, we can find more than one competent model for an image (e.g., for image (d)). To this end, we propose a novel approach that dynamically fuses multi-modal information of online images, derived through Convolutional Neural Networks (CNNs), to adequately identify the sensitive image content. In summary, we make the following contributions:

- Our significant contribution is to estimate the competence of object, scene and tag modalities for privacy prediction and dynamically identify the most competent modalities for a target image whose privacy has to be predicted.
- We derive “competence” features from the neighborhood regions of a target image and learn classifiers on them to identify whether a modality is competent to accurately predict the privacy of the target image. To derive these features, we consider privacy and visual neighborhoods of the target image to bring both sensitive and visually similar image content closer.
- We provide an in-depth analysis of our algorithm in an ablation setting, where we record the performance of the proposed approach by removing its various components. The analysis outline the crucial components of our approach.
- Our results show that we identify images’ sensitive content more accurately than single modality models (object, scene, and tag), multi-modality baselines and prior approaches of privacy prediction, depicting that the approach optimally combines the

multi-modality for privacy prediction.

5.2 Related Work

We briefly review the related work as follows.

Ensemble models and Multi-Modality. Several works used ensemble classifiers (or bagging) to improve image classifications^{195–197}. Bagging is an ensemble technique that builds a set of diverse classifiers, each trained on a random sample of the training data to improve the final (aggregated) classifier confidence^{198;199}. Dynamic ensembles that extend bagging have also been proposed^{200–202} wherein a pool of classifiers are trained on a single feature set (single modality) using the bagging technique^{198;199}, and the competence of the base classifiers is determined dynamically.

Ensemble classifiers are also used in the multi-modal setting^{203;204}, where different modalities have been coupled, e.g., images and text for image retrieval²⁰⁵ and image classification²⁰³, and audio and visual signals for speech classification²⁰⁶. Zahavy et al.²⁰⁷ highlighted that classifiers trained on different modalities can vary in their discriminative ability and urged the development of optimal unification methods to combine different classifiers. Besides, merging the Convolutional Neural Network (CNN) architectures corresponding to various modalities, that can vary in depth, width, and the optimization algorithm can become very complex. However, there is a potential to improve the performance through multi-modal information fusion, which intrigued various researchers^{205;208–210}. For example, Frome et al.²¹¹ merged an image network¹⁷ with a Skip-gram Language Model to improve classification on ImageNet. Zahavy et al.²⁰⁷ proposed a policy network for multi-modal product classification in e-commerce using text and visual content, which learns to choose between the input signals. Feichtenhofer et al.²¹² fused CNNs both spatially and temporally for activity recognition in videos to take advantage of the spatio-temporal information present in videos. Wang et al.²¹³ designed an architecture to combine object networks and scene networks, which extract useful information such as objects and scene contexts for event understanding. Co-training approaches²¹⁴ use multiple views (or modalities) to “guide” different classifiers in

the learning process. However, co-training methods are semi-supervised and assume that all views are “sufficient” for learning. In contrast with the above approaches, we aim to capture different aspects of images, obtained from multiple modalities (object, scene, and tags), with each modality having a different competence power, and perform dynamic multi-modal fusion for image privacy prediction.

Online Image Privacy. Several works are carried out to study users’ privacy concerns in social networks, privacy decisions about sharing resources, and the risk associated with them⁴⁶⁻⁵¹. Ahern et al.⁵ examined privacy decisions and considerations in mobile and online photo sharing. They explored critical aspects of privacy such as users’ consideration for privacy decisions, content and context based patterns of privacy decisions, how different users adjust their privacy decisions and user behavior towards personal information disclosure. The authors concluded that applications, which could support and influence user’s privacy decision-making process should be developed. Jones and O’Neill⁵² reinforced the role of privacy-relevant image concepts. For instance, they determined that people are more reluctant to share photos capturing social relationships than photos taken for functional purposes; certain settings such as work, bars, concerts cause users to share less. Besmer and Lipford⁵³ mentioned that users want to regain control over their shared content, but meanwhile, they feel that configuring proper privacy settings for each image is a burden. Buschek et al.³³ presented an approach to assign privacy to shared images using metadata (location, time, shot details) and visual features (faces, colors, edges). Zerr et al.⁷ developed the PicAlert dataset, containing Flickr photos, to help detect private images and also proposed a privacy-aware image classification approach to learn classifiers on these Flickr photos. Authors considered image tags and visual features such as color histograms, faces, edge-direction coherence, and SIFT for the privacy classification task. Squicciarini et al.^{12,13} found that SIFT and image tags work best for predicting sensitivity of user’s images. Given the recent success of CNNs, Tran et al.¹⁴, and Tonge and Caragea^{24,27} showed promising privacy predictions compared with visual features such as SIFT and GIST. Yu et al.⁴¹ adopted CNNs to achieve semantic image segmentation and also learned object-privacy relatedness to identify privacy-sensitive objects.

Spyromitros-Xioufis et al.⁴⁰ used features extracted from CNNs to provide personalized image privacy classification, whereas Zhong et al.³² proposed a Group-Based Personalized Model for image privacy classification in online social media sites. Despite that an individual’s sharing behavior is unique, Zhong et al.³² argued that personalized models generally require large amounts of user data to learn reliable models, and are time and space consuming to train and to store models for each user, while taking into account possible deviations due to sudden changes of users’ sharing activities and privacy preferences. Orekondy et al.³ defined a set of privacy attributes, which were first predicted from the image content and then used these attributes in combination with users preferences to estimate personalized privacy risk. The authors used official online social network rules to define the set of attributes, instead of collecting real user’s opinions about sensitive content and hence, the definition of sensitive content may not meet a user’s actual needs¹⁰⁴. Additionally, for privacy attribute prediction, the authors fine-tuned a CNN pre-trained on object dataset. In contrast, we proposed a dynamic multi-modal fusion approach to determine which aspects of images (objects, scenes or tags) are more competent to predict images’ privacy.

5.3 Multi-Modality

The sensitive content of an image can be perceived by the presence of one or more objects, the scenes described by the visual content and the description associated with it in the form of tags^{12;24;26;27}. We derive features (object, scene, tags) corresponding to the multi-modal information of online images as follows.

Object (F^o): Detecting objects from images is clearly fundamental to assessing whether an image is of private nature. For example, a single element such as a firearm, political signs, may be a strong indicator of a private image. Hence, we explore the image descriptions extracted from VGG-16¹⁹, a CNN pre-trained on the ImageNet dataset¹⁶ that has 1.2M+ images labeled with 1,000 object categories. The VGG-16 network implements a 16 layer deep network; a stack of convolutional layers with a very small receptive field: 3×3 followed by fully-connected layers. The architecture contains 13 convolutional layers and 3 fully-

connected layers. The input to the network is a fixed-size 224×224 RGB image. The activation of the fully connected layers capture the complete object contained in the region of interest. Hence, we use the activation of the last fully-connected layer of VGG-16, i.e., fc_8 as a feature vector. The dimension of object features F^o is 1000.

Scene (F^s): As consistently shown in various user-centered studies⁵, the context of an image is a potentially strong indicator of what type of message or event users are trying to share online. These scenes, e.g., some nudity, home, fashion events, concerts are also often linked with certain privacy preferences. Similar to object features, we obtain the scene descriptors derived from the last fully-connected layer of the pre-trained VGG-16¹⁷ on the Places2 dataset which contains 365 scene classes within 2.5 million images²⁵. The dimension of scene features F^s is 365.

Image Tags (F^t): For image tags, we employ the CNN architecture of Collobert et al.¹¹⁵. The network contains one convolution layer on top of word vectors obtained from an unsupervised neural language model. The first layer embeds words into the word vectors pre-trained by Le and Mikolov¹¹⁶ on 100 billion words of Google News, and are publicly available. The next layer performs convolutions on the embedded word vectors using multiple filter sizes of 3, 4 and 5, where we use 128 filters from each size and produce a tag feature representation. A max-pooling operation over the feature map is applied to capture the most important feature of length 256 for each feature map. To derive these features, we consider two types of tags: (1) user tags, and (2) deep tags. Because not all images on social networking sites have user tags or the set of user tags is very sparse²³, we predict the top d object categories (or deep tags) from the probability distribution extracted from CNN.

Object + Scene + Tag (F^{ost}): We use the combination of the object, scene, and tag features to identify the neighborhood of a target image. We explore various ways given in²¹² to combine the features. For example, we use fc_7 layer of VGG to extract features of equal length of 4096 from both object-net and scene-net and consider the max-pooling of these vectors to combine these features. Note that, in this work, we only describe the combination of features that worked best for the approach. We obtain high-level object “ F^o ” and scene “ F^s ” features from fc_8 layer of object-net and scene-net respectively and

concatenate them with the tag features as follows: $F^{ost} = f^{cat}(F^o, F^s, F^t)$. $F^{ost} = F^o(i), 1 \leq i \leq 1000, F^{ost}(i + 1000) = F^s(i), 1 \leq i \leq 365, F^{ost}(i + 1365) = F^t(i), 1 \leq i \leq 256$.

5.4 Proposed approach

We seek to classify a given image into one of the two classes: *private* or *public*, based on users’ general privacy preferences. To achieve this, we depart from previous works that use the same model on all image types (e.g., portraits, bedrooms, and legal documents), and propose an approach called “Dynamic Multi-Modal Fusion for Privacy Prediction” (or DMFP), that effectively fuses multi-modalities (object, scene, and tags) and dynamically captures different aspects or particularities from image. Specifically, the proposed approach aims to estimate the competence of models trained on these individual modalities for each target image (whose privacy has to be predicted) and dynamically identifies the subset of the most “competent” models for that image. The rationale for the proposed method is that for a particular type of sensitive content, some modalities may be important, whereas others may be irrelevant and may simply introduce noise. Instead, a smaller subset of modalities may be significant in capturing a particular type of sensitive content (e.g., objects for portraits, scenes for interior home or bedroom, and tags for legal documents, as shown in Figure 5.1).

The proposed approach considers three stages to predict the privacy of a target image, wherein we first identify the neighborhood images that are visually similar and/or have similar sensitive content as the target image (Section 5.4.1). Then, using a set of base classifiers, each trained on an individual modality, we estimate the competence of the modalities by determining which modalities classify the neighborhood images correctly (Section 5.4.2). The goal here is to select the most competent modalities for a particular type of images (e.g., scene for home images). Finally, we fuse the decisions of the most competent base classifiers (corresponding to the most competent modalities) and predict the privacy label for the target image (Section 5.4.3).

Our approach considers two datasets, denoted as \mathcal{D}^T and \mathcal{D}^E , that contain images labeled as *private* or *public*. We use the dataset \mathcal{D}^T to train a base classifier for each modality

to predict whether an image is public or private. Particularly, we train 3 base classifiers $\mathcal{B} = \{B^o, B^s, B^t\}$ on the corresponding modality feature sets from \mathcal{F} . Note that we use the combination of feature sets F^{ost} only for visual content similarity and do not train a base classifier on it. The competences of these base classifiers are estimated on the \mathcal{D}^E dataset. We explain the stages of the proposed approach as follows. The notation used is shown in Table 5.1.

Notation	Description
\mathcal{D}^T	$= \{(I_1, Y_1), \dots, (I_m, Y_m)\}$ a dataset of labeled images for base classifier training.
\mathcal{D}^E	$= \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ a dataset of labeled images for competence estimation.
T	A target image with an unknown privacy label.
\mathcal{F}	$= \{F^o, F^s, F^t, F^{ost}\}$ a collection of modality feature sets of object, scene, tag, and object+scene+tag, respectively.
\mathcal{B}	$= \{B^o, B^s, B^t\}$ a set of base classifiers trained on corresponding modality feature sets from \mathcal{F} (e.g., B^o is trained on F^o).
N_V^T	The visual similarity based neighborhood of image T estimated using visual content features F^{ost} , i.e., the set of most similar images to T based on visual content.
k_v	The size of N_V^T , where $1 \leq k_v < n$.
N_P^T	The privacy profile based neighborhood of target T , i.e., the set of most similar images to T based on images' privacy profiles.
k_p	The size of N_P^T , where $1 \leq k_p < n$.
\mathcal{C}	$= \{C^o, C^s, C^t\}$ a set of "competence" classifiers corresponding to the base classifiers from \mathcal{B} (e.g., C^o for B^o).
Φ	$= \{\phi^o, \phi^s, \phi^t\}$ a set of "competence" feature vectors for training the "competence" classifiers.

Table 5.1: *Mathematical notations.*

5.4.1 Identification of Neighborhoods

The competence of a base classifier is estimated based on a local region where the target image is located. Thus, given a target image T , we first estimate two neighborhoods for T : (1) visual similarity based (N_V^T) and (2) privacy profile based (N_P^T) neighborhoods.

The neighborhood N_V^T of target image T consists of the k_v most similar images from \mathcal{D}^E using visual content similarity. Specifically, using the F^{ost} features obtained by concatenating object, scene, and tag features (as explained in Section 5.3), we determine the k_v most visually similar images to T by applying the K-Nearest Neighbors algorithm on the \mathcal{D}^E dataset.

The neighborhood N_P^T of target image T consists of k_p most similar images to T by calculating the cosine similarity between the privacy profile of T and images from the dataset \mathcal{D}^E . We define privacy profile (denoted by \bar{T}) of image T as a vector of posterior privacy probabilities obtained by the base classifiers \mathcal{B} i.e., $\bar{T} = \bigcup_{B_i \in \mathcal{B}} \{P(Y_T = \textit{private}|T, B_i), P(Y_T = \textit{public}|T, B_i)\}$. For image (a) in Figure 5.1, $\bar{T} = [0.62, 0.38, 0.5, 0.5, 0.29, 0.71]$. We consider the privacy profile of images because images of particular image content (bedroom images or legal documents) tend to possess similar privacy probabilities with respect to the set of base classifiers \mathcal{B} . For instance, irrespective of various kinds of bedroom images, the probabilities for a private class obtained by base classifiers \mathcal{B} , would be similar. This enables us to bring sensitive content closer irrespective of their disparate visual content. Moreover, we consider two different numbers of nearest neighbors k_v and k_p to find the neighborhoods since the competence of a base classifier is dependent on the neighborhood and estimating an appropriate number of neighbors for the respective neighborhoods reduces the noise.

5.4.2 “Competence” Estimation

We now describe how we estimate the “competence” of a base classifier. For instance, for the image (a) in Figure 5.1, scene model has a higher competence than the others, and here, we capture this competence through “competence” features and “competence” classifiers. Specifically, we train a competence classifier for each base classifier that predicts if the base classifier is competent or not for a target image T . The features for learning the competence classifiers and the competence learning are described below.

Derivation of “Competence” Features

We define three different sets of “competence” features wherein each set of these features captures a different criterion to estimate the level of competence of base classifiers dynamically. The first competence feature ϕ_1 , for image T , is derived from the neighborhood N_V^T (based on visual similarity) whereas the second competence feature ϕ_2 is obtained from the neighborhood N_P^T (based on privacy profile). The third competence feature ϕ_3 captures

the level of confidence of base classifiers for predicting the privacy of the image (T) itself. We create a “competence” feature vector by concatenating all these competence features $\phi = \{\phi_1 \cup \phi_2 \cup \phi_3\}$ into a vector of length $|\phi| = k_v + k_p + 1$. We extract such competence vectors corresponding to each base classifier in \mathcal{B} (e.g., ϕ^o for B^o , refer to Figure 5.2). We extract these “competence” features as follows.

ϕ_1 : A vector of k_v entries that captures the correctness of a base classifier in the visual neighborhood region N_V^T . An entry j in ϕ_1 is 1 if a base classifier $B_i \in \mathcal{B}$ accurately predicts privacy of image $X_j \in N_V^T$, and is 0 otherwise, where $j = 1, \dots, k_v$. For the target image in Figure 5.2, $\phi_1 = \{1, 1, 0, 1, 0, 1, 1\}$, obtained by B^o .

ϕ_2 : A vector of k_p entries that captures the correctness of a base classifier in the privacy profile neighborhood region N_P^T . An entry j in ϕ_2 is 1 if a base classifier $B_i \in \mathcal{B}$ accurately predicts privacy of image $X_j \in N_P^T$, and is 0 otherwise, where $j = 1, \dots, k_p$. For the target image in Figure 5.2, $\phi_2 = \{1, 1, 1, 1, 1\}$, obtained using B^o .

ϕ_3 : We capture a degree of confidence of base classifiers for target image T . Particularly, we consider the maximum posterior probability obtained for target image T using base classifier B_i i.e. $\text{Max}(P(Y_T = \text{Private}|T, B_i), P(Y_T = \text{Public}|T, B_i))$, where $B_i \in \mathcal{B}$. For the target image in Figure 5.2, $\phi_3 = 0.67$, obtained using B^o .

“Competence” Learning

We learn the “competence” of a base classifier by training a binary “competence” classifier on the dataset \mathcal{D}^E in a Training Phase. A competence classifier predicts whether a base classifier is competent or not for a target image. Algorithm 3 describes the “competence” learning process in details. Mainly, we consider images from \mathcal{D}^E as target images (for the training purpose only) and identify both the neighborhoods (N_V, N_P) from the dataset \mathcal{D}^E itself (Alg. 3, lines 6–8). Then, we extract “competence” features for each base classifier in \mathcal{B} based on the images from these neighborhoods (Alg. 3, line 10). To reduce noise, we extract “competence” features by considering only the images belonging to both the neighborhoods. On these “competence” features, we train a collection of “competence”

Algorithm 3 The ‘‘Competence’’ Learning

```
1: Input: A dataset  $\mathcal{D}^E = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$  of labeled images;  $F_{X_j}^{ost}$  combination of
   modality feature sets for  $X_j$ ; a set of base classifiers  $\mathcal{B} = \{B^o, B^s, B^t\}$ .
2: Output: A set of ‘‘competence’’ classifiers  $\mathcal{C} = \{C^o, C^s, C^t\}$ .
3:  $\mathcal{D} = \{D^o, D^s, D^t\} \leftarrow \emptyset$ ; // Datasets for training competence classifiers, initially empty.
4:  $\mathcal{C} \leftarrow \emptyset$ ; // A set of competence classifiers, initially empty.
5: for all  $X_j \in \mathcal{D}^E$  do
6:    $N_V^{X_j} \leftarrow \text{IdentifyVisualNeighborhood}(k_v, X_j, F_{X_j}^{ost}, \mathcal{D}^E)$ ; //  $k_v$  nearest neighbors of  $X_j$ 
   obtained using visual content similarity.
7:    $\overline{X_j} \leftarrow \text{ComputePrivacyProfile}(X_j, \mathcal{B})$ ; // Privacy profile.
8:    $N_P^{X_j} = \text{IdentifyPrivacyNeighborhood}(k_p, \overline{X_j}, \overline{\mathcal{D}^E})$ ; //  $k_p$  most similar images of  $\overline{X_j}$ 
   obtained using privacy profile similarity.
9:   for all  $B_i \in \mathcal{B}$  do // Iterate through the set of base classifiers.
10:     $\phi_{i,j} \leftarrow \text{CompetenceFeatures}(X_j, N_V^{X_j}, N_P^{X_j}, B_i)$ ;
11:    if  $\text{Predict}(B_i, X_j) = Y_j$  then // predicted correctly.
12:       $L_{i,j} \leftarrow 1$ ; //  $B_i$  is competent for  $X_j$ .
13:    else
14:       $L_{i,j} \leftarrow 0$ ; //  $B_i$  is not competent for  $X_j$ .
15:    end if
16:     $D^i \leftarrow D^i \cup \{(\phi_{i,j}, L_{i,j})\}$ 
17:  end for
18: end for
19: for all  $D^i \in \mathcal{D}$  do // Train competence classifiers.
20:    $C_i \leftarrow \text{TrainCompetenceClassifier}(D^i)$ ;
21:    $\mathcal{C} \leftarrow \mathcal{C} \cup C_i$ 
22: end for
23: return  $\mathcal{C}$ ;
```

classifiers \mathcal{C} corresponding to each base classifier in \mathcal{B} (Alg. 3, lines 19–22). Precisely, we train 3 competence classifiers $\mathcal{C} = \{C^o, C^s, C^t\}$. To train ‘‘competence’’ classifier $C_i \in \mathcal{C}$, we consider label $L_i = 1$ if base classifier $B_i \in \mathcal{B}$ predicts the correct privacy of a target image (here, $X_j \in \mathcal{D}^E$), otherwise 0 (Alg. 3, lines 11–16).

5.4.3 Dynamic Fusion of Multi-Modality

In this stage, for given target image T , we dynamically determine the subset of most competent base classifiers. We formalize the process of base classifier selection in Algorithm 4. The algorithm first checks the agreement on the privacy label between all the base classifiers in \mathcal{B} (Alg. 4, line 5). If not all the base classifiers agree, then we estimate the competence of all

Algorithm 4 Dynamic Fusion of Multi-Modality

```
1: Input: A target image  $T$ ;  $\mathcal{D}^E = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$  a dataset of labeled images;
    $F_T^{ost}$  combination of modality feature sets for  $T$ ; a set of base classifiers  $\mathcal{B} = \{B^o, B^s, B^t\}$ ;
   and a set of competence classifiers  $\mathcal{C} = \{C^o, C^s, C^t\}$ .
2: Output: Privacy label  $Y_T$ .
3:  $B' \leftarrow \emptyset$ ; // the subset of most competent base classifiers.
4:  $CS \leftarrow \emptyset$ ; // the set of competence scores.
5:  $T_{ba} \leftarrow \text{Agreement}(\mathcal{B}, T)$ ; // Base classifiers' agreement on  $T$ 's label.
6: if  $T_{ba} \leq |\mathcal{B}|$  then
7:    $N_V^T \leftarrow \text{IdentifyVisualNeighborhood}(k_v, T, F_T^{ost}, \mathcal{D}^E)$ ; //  $k_v$  nearest neighbors of  $T$ 
   obtained using visual content similarity.
8:    $\bar{T} \leftarrow \text{ComputePrivacyProfile}(T, \mathcal{B})$ ; // Privacy profile of  $T$ .
9:    $N_P^T = \text{IdentifyPrivacyNeighborhood}(k_p, \bar{T}, \mathcal{D}^E)$ ; //  $k_p$  most similar images of  $\bar{T}$  ob-
   tained using privacy profile similarity.
10:  for all  $B_i \in \mathcal{B}$  &  $C_i \in \mathcal{C}$  do // Iterate through the set of base and competence classifiers.
11:     $\phi_i \leftarrow \text{CompetenceFeatures}(T, N_V^T, N_P^T, B_i)$ ;
12:     $CS_i \leftarrow \text{PredictCompetence}(F_i, C_i)$ ; // Predict competence score for base classifier
     $B_i$ .
13:    if  $CS_i > 0.5$  then // If the predicted competence score is greater than 0.5 then the
    base classifier  $B_i$  is competent.
14:       $B' \leftarrow B' \cup \{B_i\}$ 
15:       $CS \leftarrow CS \cup \{CS_i\}$ 
16:    end if
17:  end for
18:   $Y_T = \text{WeightedMajorityVote}(T, B', CS)$  // Votes are first weighted by the competence
    score and then majority vote is taken.
19: end if
20: return  $Y_T$ 
```

the base classifiers and identify the subset of most competent base classifiers for the target image as follows. Given target image T , Algorithm 4 first identifies both the neighborhoods (N_V^T, N_P^T) using the visual features F^{ost} and privacy profile from \mathcal{D}^E dataset (Alg. 4, lines 7–9). Using these neighborhoods, we extract “competence” feature vectors (explained in Section 5.4.2) and provide them to the respective “competence” classifiers in \mathcal{C} (learned in the Training Phase) to predict competence score of base classifier B_i . If the competence score is greater than 0.5, then base classifier B_i is identified as competent to predict the privacy of target image T (Alg. 4, lines 10–17). Finally, we weight votes of the privacy labels predicted by the subset of most competent base classifiers by their respective “competence” score and take a majority vote to obtain the final privacy label for target image T (Alg. 4, line 18).

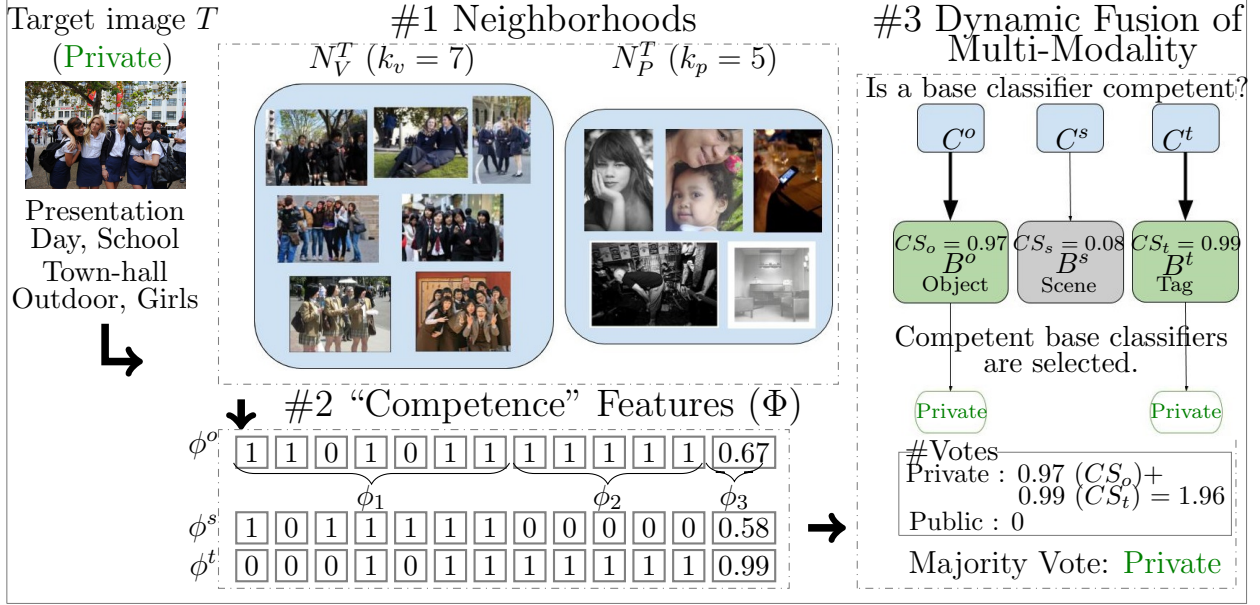


Figure 5.2: Illustration of the proposed approach using an example.

A “competence” score CS_i is given as a probability of base classifier B_i being competent. We consider the majority vote of the most competent base classifiers because certain images (e.g., vacation) might require more than one base classifiers (object and scene) to predict the appropriate privacy. If both the privacy classes (private and public) get the same number of votes, then the class of a highest posterior probability is selected.

Illustration of the Proposed Approach

Figure 5.2 shows the illustration of the proposed approach through an anecdotal example. We consider a target image T whose privacy has to be predicted. For T , we first identify two neighborhoods: (1) visual content (N_V^T), 2. privacy profile (N_P^T). For N_V^T , we use visual content features F^{ost} to compute the similarity between target image T and the images from the dataset \mathcal{D}^E . The top $k_v = 7$ similar images for N_V^T are shown in the figure (left blue rectangle). Likewise, for N_P^T , we compute the similarity between privacy profile of the target image \bar{T} and privacy profiles of images in \mathcal{D}^E . We show the top $k_p = 5$ similar images for N_P^T in the right blue rectangle of Figure 5.2. From these neighborhoods, we derive a “competence” feature vector ϕ for each base classifier in \mathcal{B} (e.g., ϕ^o for B^o). We show these

“competence” features in the figure as a matrix of feature values. We input these features to the respective “competence” classifiers from \mathcal{C} (e.g., ϕ^o to C^o), that predict whether a base classifier $B_i \in B$ is competent to predict correct privacy label of the target image (T). The “competence” classifiers (C^o, C^s, C^t) are shown as blue rectangles on the right side of Figure 5.2. The base classifiers B^o and B^t are predicted as competent and hence are selected to obtain the final privacy label for the target image. The competent base classifiers are shown in green rectangles on the right side of Figure 5.2. Once we select the competent base classifiers, we take a weighted majority vote on the privacy labels, predicted by these base classifiers. For example, in this case, the competent base classifiers B^o and B^t predict the privacy of T as “private,” and hence, the final privacy label of T is selected as “private.” It is interesting to note that the target image (T) contains “outdoor” scene context that is not useful to predict the correct privacy label and hence, the scene model B^s is not selected by the proposed approach for the target image.

5.5 Dataset

We evaluate our approach on a subset of 32,000 Flickr images sampled from the PicAlert dataset, made available by Zerr et al. ⁷. PicAlert consists of Flickr images on various subjects, which are manually labeled as *public* or *private* by external viewers. The guideline to select the label is given as: private images belong to the private sphere (like self-portraits, family, friends, someone’s home) or contain information that one would not share with everyone else (such as private documents). The remaining images are labeled as public. The dataset of 32,000 images is split in \mathcal{D}^T , \mathcal{D}^E and *Test* sets of 15,000, 10,000 and 7,000 images, respectively. Each experiment is repeated 5 times with a different split of the three subsets (obtained using 5 different random seeds) and the results are averaged across the five runs. The public and private images are in the ratio of 3:1 in all subsets.

5.6 Experiments and Results

We evaluate the privacy prediction performance obtained using the proposed approach DMFP, where we train a set of base classifiers \mathcal{B} on images in the dataset \mathcal{D}^T , and dynamically estimate the “competence” of these base classifiers for target images in *Test* by identifying neighborhoods (N_V, N_P) using images in \mathcal{D}^E . We first consider various values of neighborhood parameters k_v and k_p and show their impact on the performance of the proposed approach. Then, we compare the performance of the proposed approach with the performance obtained using three types of mechanisms: (1) components of the proposed approach, that are used to fuse the multi-modal characteristics of online images, (2) the state-of-the-art approaches for privacy prediction, and (3) strong baselines that select models based on their competence (e.g., Zahavy et al. ²⁰⁷) and that attempt to yield the optimal combination of base classifiers (for instance, using stacked ensemble classifiers).

Evaluation Setting. We train base classifiers (\mathcal{B}) using the Calibrated linear Support Vector Machine (SVM) implemented in Scikit-learn library^{215;216} to predict more accurate probabilities. We use 3-fold cross-validation on the dataset \mathcal{D}^T to fit the linear SVM on the 2-folds, and the remaining fold is used for calibration. The probabilities for each of the folds are then averaged for prediction. We train “competence” classifiers (\mathcal{C}) on the dataset \mathcal{D}^E using logistic regression to predict “competence” scores between 0 – 1 for base classifiers. If base classifier B_i gets a “competence” score greater than 0.5 then the base classifier is considered competent. To derive features from CNN, we use pre-trained models presented by the VGG-16 team in the ILSVRC-2014 competition¹⁹ and the CAFFE framework⁴². For deep tags, we consider top $d = 10$ object labels as $d = 10$ worked best.

Exploratory Analysis.

We provide exploratory analysis in Table 5.2 to highlight the potential of merging object, scene and tag modality for privacy prediction. We predict privacy for images in the *Test* set using base classifiers in \mathcal{B} and obtain “private” (Pr), “public” (Pu) and “overall” (O) accuracy for: (a) a modality is correct (e.g., object), (b) all modalities are correct, (c) all

Test	Pr(%)	Pu(%)	O(%)
Object is correct	49	95.7	84.8
Scene is correct	51	94.7	84.4
Tag is correct	57	91.1	83
All are correct	30	87.3	73.9
All are wrong	27	1.5	7.4
Atleast one modality is correct	73	98.5	92.6

Table 5.2: *Exploratory analysis.*

modalities are wrong, and (d) at least one modality is correct. Table 5.2 shows that out of the three base classifiers (top 3 rows), the tag model yields the best accuracy for the private class (57%). Interestingly, the results for “at least one modality is correct” (73%) show that using multi-modality, there is a huge potential (16%) to improve the performance of the private class. This large gap is a promising result for developing multi-modality approaches for privacy prediction. Next, we evaluate DMFP that achieved the best boost in the performance for the private class using these modalities.

5.6.1 Impact of Parameters k_v and k_p on DMFP

We show the impact of neighborhood parameters, i.e., k_v and k_p on the privacy prediction performance obtained by the proposed approach DMFP. k_v and k_p are used to identify visual (N_V) and privacy profile (N_P) neighborhoods of a target image, respectively (Alg. 4 lines 7–8). We experiment with a range of values for both the parameters $k_v, k_p = \{10, 20, \dots, 100, 200, \dots, 1000\}$, in steps of 10 upto 100 and then in steps of 100. We also experiment with larger k_v and k_p values, but for better visualization, we only show the values with significant results. Figure 5.3 shows the F1-measure obtained (using 3-fold cross-validation on the \mathcal{D}^E dataset) for the private class for various k_v and k_p values. We notice that when we increase the k_v parameter the performance increase whereas when we increase k_p parameter, the performance increases upto $k_p = 200$, then the performance decreases gradually. The results show that the performance is quite sensitive to changes in the privacy neighborhood (N_P) parameter k_p , but relatively insensitive to changes in the visual neighborhood (N_V) parameter k_v . We get the best performance for $k_v = 900$ and $k_p = 100$.

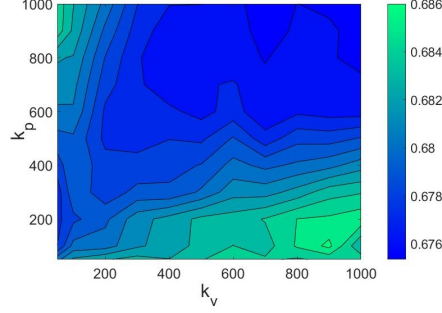


Figure 5.3: *F1-measure obtained for various k_v and k_p values.*

Features	Private			Public			Overall			
	Pre.	Re.	F1	Pre.	Re.	F1	Acc. (%)	Pre.	Re.	F1
DMFP	0.752	0.627	0.684	0.891	0.936	0.913	86.36	0.856	0.859	0.856
Components of the proposed approach										
DMFP- N_V	0.763	0.575	0.655	0.879	0.945	0.91	85.79	0.85	0.852	0.847
DMFP- N_P	0.74	0.572	0.645	0.877	0.938	0.907	85.21	0.843	0.847	0.841
$N_V - CL$	0.79	0.534	0.637	0.87	0.956	0.911	85.71	0.85	0.851	0.843
$N_P - CL$	0.788	0.537	0.639	0.87	0.956	0.911	85.73	0.85	0.851	0.843
$\{N_V + N_P\} - CL$	0.79	0.534	0.637	0.87	0.956	0.911	85.71	0.85	0.851	0.843
“Competence” Features										
DMFP- ϕ_1	0.777	0.553	0.646	0.874	0.951	0.911	85.74	0.849	0.852	0.844
DMFP- ϕ_2	0.74	0.565	0.641	0.875	0.939	0.906	85.11	0.842	0.846	0.84
DMFP- ϕ_3	0.752	0.627	0.683	0.891	0.936	0.913	86.35	0.856	0.859	0.856

Table 5.3: *Evaluation of dynamic multi-modal fusion for privacy prediction (DMFP).*

We use these parameter values in the next experiments.

5.6.2 Evaluation of the Proposed Approach

We evaluate the proposed approach DMFP for privacy prediction in an ablation experiment setting. Specifically, we remove a particular component of the proposed approach DMFP and compare the performance of DMFP before and after the removal of that component. We consider excluding several components from DMFP: (1) the visual neighborhood N_V (DMFP- N_V), (2) the privacy profile neighborhood N_P (DMFP- N_P), (3) “competence” features (e.g., DMFP- ϕ_1), and (4) base classifier selection without “competence” learning (e.g., $N_V - CL$). For option (4), we consider a simpler version of the proposed algorithm,

in which we do not learn a competence classifier for a base classifier; instead, we rely solely on the number of accurate predictions of the samples from a neighborhood. We evaluate it using images from three regions: (a) neighborhood N_V only ($N_V - CL$), (b) neighborhood N_P only ($N_P - CL$), and (c) both the neighborhoods N_P and N_V ($\{N_P + N_V\} - CL$).

Table 5.3 shows the class-specific (private and public) and overall performance obtained by the proposed approach (DMFP) and after removal of its various components detailed above. Primarily, we wish to identify whether the proposed approach characterizes the private class effectively as sharing private images on the Web with everyone is not desirable. We observe that the proposed approach achieves the highest recall of 0.627 and F1-score of 0.684 (private class), which is better than the performance obtained by eliminating the essential components (e.g., neighborhoods) of the proposed approach. We notice that if we remove either of the neighborhood N_V or N_P , the recall and F1-score drop by 5% and 4%. This suggests that both the neighborhoods (N_V, N_P) are required to identify an appropriate local region surrounding a target image. It is also interesting to note that the performance of DMFP- N_P (removal of N_P) is slightly lower than the performance of DMFP- N_V (removal of N_V), depicting that neighborhood N_P is helping more to identify the competent base classifier(s) for a target image. The N_P neighborhood brings images closer based on their privacy probabilities and hence, is useful to identify the competent base classifier(s) (this is evident in Figure 5.2). We also show that when we remove competence learning (CL) i.e., $N_V - CL$, $N_P - CL$, and $\{N_V + N_P\} - CL$, the precision improves by 4% (private class), but the recall and F1-score (private class) drops by 9% and 5% respectively, showing that competence learning is necessary to achieve the best performance.

We also remove the “competence” features one by one and record the performance of DMFP to understand which competence features are essential. Table 5.3 shows that when we remove feature ϕ_1 corresponding to the neighborhood N_V , the performance drops significantly ($\approx 4\%$). Likewise, when we remove ϕ_2 (feature corresponding the N_P region), we notice a similar decrease of 4% in the F1-score of private class. Note that, when we remove the “competence” features corresponding to their neighborhoods (such as ϕ_1 for N_V and ϕ_2 for N_P), we get nearly similar performance as we remove the respective neighborhoods

Features	Private			Public			Overall			
	Pre.	Re.	F1	Pre.	Re.	F1	Acc. (%)	Pre.	Re.	F1
DMFP	0.752	0.627	0.684	0.891	0.936	0.913	86.36	0.856	0.859	0.856
Object (B^o)	0.772	0.513	0.616	0.864	0.953	0.907	84.99	0.843	0.85	0.838
Scene (B^s)	0.749	0.51	0.606	0.863	0.947	0.903	84.45	0.836	0.844	0.833
Image tags (B^t)	0.662	0.57	0.612	0.873	0.91	0.891	83.03	0.823	0.83	0.826

Table 5.4: *Dynamic multi-modal fusion for privacy prediction (DMFP) vs. base classifiers of DMFP.*

from the proposed approach (DMFP- N_V and DMFP- N_P); implying that removing “competence” features (e.g., ϕ_1) is as good as removing the corresponding neighborhood (N_V). However, a close look at the performance suggests that the performance obtained using DMFP- ϕ_1 (recall of 0.553) is slightly worse than the performance of DMFP- N_V (recall of 0.57). Similarly, for DMFP- ϕ_2 , the performance (recall) decrease from 0.572 obtained using DMFP- N_P to 0.565. The performance decrease can be explained as when we remove the neighborhood N_V or N_P , the respective “competence” features are empty, and that might be helpful for some cases (as zero-valued ϕ_2 feature of ϕ^s was helpful in Figure 5.2). Additionally, the recall of DMFP- N_V and DMFP- N_P are similar whereas the recall of DMFP- ϕ_1 (0.553) is slightly worse than the recall of DMFP- ϕ_2 (0.565). The results suggests that the neighborhood N_V is more dependent on the “competence” features as compared to the neighborhood N_P . We experimented with the probability based “competence” features (instead of boolean features), but did not yield improvements in the performance.

5.6.3 Proposed Approach vs. Base Classifiers

We compare privacy prediction performance obtained by the proposed approach DMFP with the set of base classifiers \mathcal{B} : 1. object (B^o), 2. scene (B^s), and 3. image tags (B^t).

Table 5.4 compares the performance obtained by the proposed approach (DMFP) and





Model				
	(a)	(b)	(c)	(d)
DMFP	✓	✓	✓	✗
Object	✗	✓	✓	✗
Scene	✓	✗	✓	✗
Tags	✓	✓	✗	✗

Figure 5.4: *Predictions for private images.*

\mathcal{B}	overall	private	public
object	16.52	14.00	23.75
scene	27.71	21.78	42.63
Tags	37.02	26.90	58.79

Table 5.5: *Errors corrected (%)*.

base classifiers. We achieve the highest performance as compared to the base classifiers and show a maximum improvement of $\approx 10\%$ in the F1-score of private class. We notice that our approach based on multi-modality yields an improvement of 11% over the recall of object and scene models and an improvement of $\approx 6\%$ over the recall of the tag model, that is the best-performing single modality model obtained for the private class from the exploratory analysis (refer to Table 5.2). Still, our approach makes some errors (See Table 5.2 and 5.3, 73% vs. 62%). A close look at the errors discovered that a slight subjectivity of annotators could obtain different labels for similar image subjects (e.g., food images are very subjective).

Error Analysis

We perform error analysis to further analyze the results of the proposed approach. We first determine the errors generated by all the base classifiers in \mathcal{B} and corrected by the proposed approach DMFP. We calculate the percentage of corrected errors for private class, public class and overall (considering both the classes) and show them in Table 5.5. We compute the percentage of corrected errors as the number of corrected errors of private (or public) class over the total number of private (or public) class errors. We calculate the fraction of overall corrected errors by considering both public and private classes. The table shows that we correct 14% – 27% of private class errors, 18% – 58% of public class errors and overall we

eliminate 16% – 37% errors. Note that errors generated for the private class are much larger than the public class (See Table 5.4) and hence, even a comparatively smaller percentage of corrected errors constitute to a significant improvement. We also analyze results by showing predictions of samples in Figure 5.4, for which at least one base classifier fails to predict the correct privacy of an image. For instance, for example (b), scene model failed to predict the correct privacy of the image; however, DMFP identifies the competent base classifiers, i.e., object, and tag and predict the correct privacy label. We also show an example (image (d)) for which all the base classifiers fail to predict the correct privacy class and hence, the proposed approach also fails to predict the correct label. The image of a food is very subjective and hence, generic base classifiers will not be sufficient to predict the correct labels of such images. In the future, these generic models can be extended to develop hybrid approaches, that consider both generic and subjective privacy notions to predict personalized privacy labels.

5.6.4 Proposed Approach vs. Baselines

We compare the performance of the proposed approach DMFP with multi-modality based baselines described below.

- 1. Model Selection by Zahavy et al.²⁰⁷:** The authors proposed a deep multi-modal architecture for product classification in e-commerce, wherein they learn a decision-level fusion policy to choose between image and text CNN for an input product. Specifically, the authors provide class probabilities of a product as input to the policy trained on a validation dataset and use it to predict whether image CNN (or text CNN) should be selected for the input. In other words, policy determines the competence of the CNNs for its input and thus, we consider it as our baseline. For a fair comparison, we learn 3 policies (corresponding to the competence classifiers \mathcal{C}), wherein each policy (say object policy) predicts whether the respective base classifier (object) should be selected for a target image. Note that we learn these policies on the \mathcal{D}^E dataset. Finally, we take a majority vote of the privacy label predicted by the selected base classifiers (identified by the policies) for a target image.

2. Majority Vote: We consider a majority vote as another baseline, as we use it for final selection of privacy label for a target image. Unlike our approach, a vote is taken without any pre-selection of base classifiers. We predict privacy of a target image using base classifiers in \mathcal{B} and select a label having highest number of votes.

3. Decision-level Fusion: Fixed rules, that average the predictions of the different CNNs¹⁷ or select the CNN with the highest confidence²⁰⁴. The first rule is equivalent to the majority vote baseline, and hence, we show the results for the second rule only. The second rule is given as: $Y_T = \operatorname{argmax}_i([P_i^o + P_i^s + P_i^t]/3)$, where $i = 0$ (public), 1 (private). P^o, P^s , and P^t denotes the posterior probabilities (private or public) obtained using object (B^o), scene (B^s) and tag (B^t) modality respectively.

4. Stacked Ensemble (Stacked-en): Stacking learns a meta-classifier to find an optimal combination of the base learners^{217;218}. Unlike bagging and boosting, stacking ensembles robust and diverse set of base classifiers together, and hence, we consider it as one of the baselines. We use the same set of base classifiers \mathcal{B} to encode images in \mathcal{D}^T using posterior probabilities $P(Y_I = \textit{private}|I, B_i)$ and $P(Y_I = \textit{public}|I, B_i)$ where $B_i \in \mathcal{B}$. We train a meta-classifier on this encoded \mathcal{D}^T dataset using calibrated SVM classifier. We use this meta-classifier to predict privacy class of an encoded target image T (using the posterior probabilities obtained by the base classifiers $P(Y_T = \textit{private}|T, B_i)$ and $P(Y_T = \textit{public}|T, B_i)$). As we use \mathcal{D}^E only to learn ‘‘competence’’ classifiers, we do not consider it for training a meta-classifier for a fair comparison.

5. Clusters-based Models (Cluster-en): We create 5 clusters of \mathcal{D}^T dataset using hierarchical clustering mechanism and the combination of object, scene and tag features (F^{ost}). We train a calibrated SVM model on each cluster using the combination of features F^{ost} . For target image T , the most relevant cluster is identified using $k = 15$ nearest neighbors, and the model trained on that cluster is used to predict the privacy of the target image. We consider this as another baseline, because clustering images that are shared online, brings similar image types (e.g., portraits) together and models trained on these clusters can be competent to predict privacy of target images of respective image types. The number of clusters and neighbors are estimated based on the \mathcal{D}^E dataset.

Features	Private			Public			Overall			
	Pre.	Re.	F1	Pre.	Re.	F1	Acc. (%)	Pre.	Re.	F1
DMFP	0.752	0.627	0.684	0.891	0.936	0.913	86.36	0.856	0.859	0.856
Zahavy et al. ²⁰⁷	0.662	0.568	0.612	0.873	0.911	0.891	83.02	0.82	0.825	0.821
Majority Vote	0.79	0.534	0.637	0.87	0.956	0.911	85.71	0.85	0.851	0.843
Decision-level Fusion	0.784	0.555	0.65	0.874	0.953	0.912	85.94	0.852	0.853	0.846
Stacked-En	0.681	0.59	0.632	0.879	0.915	0.897	83.86	0.829	0.834	0.831
Cluster-En	0.748	0.429	0.545	0.845	0.956	0.897	83.17	0.822	0.831	0.814

Table 5.6: *Dynamic multi-modal fusion for privacy prediction (DMFP) vs. baselines.*

Table 5.6 compares the performance obtained by the proposed approach DMFP with the performance obtained using the baseline models. We observe that DMFP learns better privacy characteristics than baselines with respect to private class by providing improvements of 4.5% – 14% and 4% – 20% in the F1-score and recall of private class. When we learn the “competence” of the base classifiers (\mathcal{B}) on the \mathcal{D}^E dataset without identifying the neighborhoods (the first baseline, Zahavy et al.²⁰⁷), the precision, recall and F1-score drop by 9%, $\approx 6\%$, $\approx 7\%$. It is interesting to note that the precision of DMFP-CL (Refer Table 5.3, $N_V - CL$, $N_P - CL$, $\{N_V + N_P\} - CL$), i.e., 0.79 is better than the first baseline (Zahavy et al.²⁰⁷), i.e., 0.662 whereas the recall of the first baseline (0.568) is better than DMFP-CL (0.534). However, when we combine the neighborhoods ($\{N_V + N_P\}$) and the first baseline (competence learning), i.e., the proposed approach DMFP, we get better performance than each of these methods. Another detail to note that the performance of the first baseline (Zahavy et al.²⁰⁷) is very close to the image tags model (see Table 5.6, 5.4) and even though the baseline uses multi-modality, the performance does not exceed significantly over the individual base classifiers (object, scene, image). Zahavy et al.²⁰⁷ performed well for product classification, but it failed to yield improved results for privacy prediction because unlike product images or ImageNet images (that contains single object in the image), images

that are shared online are much more complex (containing multiple objects, and scene) and diverse (having different image subjects such as self-portraits, personal events). The results suggest that it is hard to generalize the competency of base classifiers on all types of image subjects and hence, the competence of the base classifiers needs to be determined dynamically based on the neighborhoods of a target image.

Table 5.6 also shows that F1-measure of private class improves from 0.637 achieved by majority vote (the second baseline), 0.65 obtained by decision-level fusion (the third baseline), 0.636 obtained by stacked-en (the fourth baseline), and 0.545 obtained by cluster-en (the fifth baseline) to 0.684 obtained by DMFP. Additionally, we notice that the proposed approach is able to achieve a performance higher than 85% in terms of all compared measures. Note that a naive baseline which classifies every image as “public” obtains an accuracy of 75%. With a paired T-test, the improvements over the baseline approaches for F1-measure of a private class are statistically significant for p-values < 0.05 .

5.6.5 Proposed Approach vs. Prior Image Privacy Prediction Works

We compare the privacy prediction performance obtained by the proposed approach DMFP with the state-of-the-art works of privacy prediction: **1. object**^{24;27} (B^o), **2. scene**²⁶ (B^s), **3. image tags**^{12;24;27} (B^t), 4. PCNH privacy framework¹⁴, and 5. Concatenation of all features²⁶. Note that the first three works are the feature sets of DMFP and are evaluated in the Experiment 5.6.3. We describe the remaining prior works (i.e., 4 and 5) in what follows.

4. PCNH privacy framework¹⁴: The framework combines features obtained from two architectures: one that extracts convolutional features (size = 24), and another that extracts object features (size = 24). The object CNN is a very deep network of 11 layers obtained by appending three fully-connected layers of size 512, 512, 24 at the end of the fully-connected layer of AlexNet¹⁷. The PCNH framework is first trained on the ImageNet dataset¹⁶ and then fine-tuned on a privacy dataset.

5. Combination of Object, Scene and User Tags (Concat)²⁶: Tonge et al.²⁶ combined object and scene tags with user tags and achieved

Features	Private			Public			Overall			
	Pre.	Re.	F1	Pre.	Re.	F1	Acc. (%)	Pre.	Re.	F1
DMFP	0.752	0.627	0.684	0.891	0.936	0.913	86.36	0.856	0.859	0.856
PCNH ¹⁴	0.689	0.514	0.589	0.862	0.929	0.894	83.15	0.819	0.825	0.818
Concat ²⁶ (F^{ost})	0.671	0.551	0.605	0.869	0.917	0.892	83.09	0.82	0.826	0.821

Table 5.7: *Dynamic multi-modal fusion for privacy prediction (DMFP) vs. prior image privacy prediction works.*

an improved performance over the individual sets of tags. Thus, we compare the proposed approach with the SVM models trained on the combination of all feature sets (F^{ost}) to show that it will not be adequate to predict the privacy of an image accurately. In our case, we consider object and scene visual features instead of tags and combine them with user tags to study multi-modality with the concatenation of visual and tag features.

Table 5.7 compares the performance obtained by the proposed approach (DMFP) and prior works. We achieve the highest performance as compared to the prior works and show a maximum improvement of $\approx 10\%$ in the F1-score of private class. We notice that our approach based on multi-modality yields an improvement of 11% over the recall of almost all the prior works (Refer Table 5.4 and 5.7). Particularly, we show improvements in terms of all measures over the PCNH framework, that uses two kinds of features object and convolutional. We found that adding high-level descriptive features such as scene context and image tags to the object features help improve the performance. In addition to the individual feature sets, we also outperform the concatenation of these feature sets (denoted as ‘Concat’), showing that ‘Concat’ could not yield an optimal combination of multi-modality. We notice that the performance of ‘Concat’ is slightly lower than the performance of base classifiers (Refer Tables 5.4 and 5.7). We find this is consistent with Zahavy et al.²⁰⁷ results, that concatenated various layers of image and tag CNN and trained the fused CNN end-to-end but did not yield a better performance than the individual CNN (image or tag).

5.7 Chapter Summary and Future Directions

In this chapter, we estimate the competence of object, scene and image tag modalities, derived through convolutional neural networks and dynamically identify the set of most competent modalities for a target image to adequately predict the class of the image as *private* or *public*. The proposed approach contains three stages wherein we first identify neighborhoods for a target image based on visual content similarity and privacy profile similarity. Then, we derive “competence” features from these neighborhoods and provide them to the “competence” classifiers to predict whether a modality is competent for the target image. Lastly, we select the subset of the most competent modalities and take a majority vote to predict privacy class of the target image. Experimental results show that our approach predicts the sensitive (or private) content more accurately than the models trained on an individual modality (object, scene, and tags), multi-modality baselines and prior privacy prediction approaches. Also, our approach could aid other applications such as event understanding, image classification, to on the fly decide which CNN (object, scene or tag) to use based on a target image.

In the future, it will be interesting to study dynamic multi-modal fusion in personalized privacy setting. Also, other types of competence learning approaches and competence features can be developed for estimating the competence of base classifiers.

Chapter 6

Summary and Discussion

In this chapter, we summarize the the contributions of this work and present future direction in our research.

6.1 Dissertation Summary

Images today are increasingly shared online on social networking sites such as Facebook, Flickr, Foursquare, and Instagram. Image sharing occurs not only within a group of friends but also more and more outside a user’s social circles for purposes of social discovery. Despite that current social networking sites allow users to change their privacy preferences, this is often a cumbersome task for the vast majority of users on the Web, who face difficulties in assigning and managing privacy settings. When these privacy settings are used inappropriately, online image sharing can potentially lead to unwanted disclosures and privacy violations. Thus, automatically predicting images’ privacy to warn users about private or sensitive content before uploading these images on social networking sites has become a necessity in our current interconnected world.

The result of our classification task is expected to aid other very practical applications. For example, a law enforcement agent who needs to review digital evidence on a suspected equipment to detect sensitive content in images and videos, e.g., child pornography. The

learning models developed here can be used to filter or narrow down the number of images and videos having sensitive or private content before other more sophisticated approaches can be applied to the data. Consider another example, images today are often stored in the cloud (e.g., Dropbox or iCloud) as a form of file backup to prevent their loss from physical damages and they are vulnerable to unwanted exposure when the storage provider is compromised. Our work can alert users before uploading their private (or sensitive) images to the cloud systems to control the amount of personal information (eg. social security number) shared through images.

Through this study, we first derive image tags, and visual content features by leveraging CNN architectures, which are used in conjunction with machine learning classifiers and then dynamically fuse these modalities to identify sensitive content accurately. Specifically, we divided this research into four tasks. The following is a summary of this dissertation:

- **The Use of Deep Features for Online Image Sharing:** In this work, we explored AI technology, i.e., deep features extracted from various CNN layers, for image privacy classification. Our results show that the deep visual features corresponding to the fully-connected layers of the AlexNet CNN outperform those corresponding to the “prob” layer. We also examined user annotated tags and deep tags (generated from the “prob” layer) and found that the combination of both the tags outperforms individual sets of tags. In addition, models trained on deep features yield improvement in performance over several baselines.
- **DeepPrivate Features For Image Privacy Prediction:** In this chapter, we provide a comprehensive study of the deep features derived from various CNN architectures of increasing depth to discover the best features that can provide an accurate privacy prediction for online images. Specifically, we explored features obtained from various layers of the pre-trained CNNs such as AlexNet, GoogLeNet, VGG-16, and ResNet and used them with SVM classifiers to predict an image’s privacy as *private* or *public*. We also fine-tuned these architectures on a privacy dataset. The study reveals that the SVM models trained on features derived from ResNet perform better than the models

trained on the features derived from AlexNet, GoogLeNet, and VGG-16. We found that the overall performance obtained using models trained on the features derived through pre-trained networks is comparable to the fine-tuned architectures. However, fine-tuned networks provide improved performance for the private class as compared to the models trained on pre-trained features. The results show remarkable improvements in the performance of image privacy prediction as compared to the models trained on CNN-based and traditional baseline features. Additionally, models trained on the deep features outperform rule-based models that classify images as private if they contain people. We also investigate the combination of user tags and deep tags derived from CNN architectures in two settings: (1) using SVM on the bag-of-tags features; and (2) applying the text CNN over these tags. We thoroughly compare these models with the models trained on the highest performing visual features obtained for privacy prediction. We further provide a detailed analysis of tags that gives insights for the most informative tags for privacy predictions. We finally show that the combination of deep visual features with these informative tags yields improvement in the performance over the individual sets of features (visual and tag).

- **Privacy-Aware Tag Recommendation for Image Sharing.** We proposed an approach to recommending privacy-aware image tags that can improve the original set of user tags and, at the same time, preserve images' privacy to help reduce the private content from the search results. Our approach draws ideas from collaborative filtering (CF). Although the user-input tags are prone to noise, we were able to integrate them in our approach and recommend accurate tags. More importantly, we simulated the recommendation strategy for newly-posted images, which had no tags attached. This is a particularly challenging problem, as in many CF approaches, the absence of items (tags in our case) may theoretically hinder the recommendations to be produced, due to the lack of enough information available to find similar images to a target image. Through our experiments, we showed that we achieve better performance for image privacy prediction with recommended tags than the original set of user tags, which in

turn indicates that the suggested tags comply to the images’ privacy. We also show that improving user tags with a set of privacy-aware recommended tags can reduce the performance gap between the tag and visual features for privacy prediction. Visual features and tag features can complement each other, and hence, can be combined to obtain improved privacy prediction performance in the future. Last, we conducted a user evaluation to inspect the quality of our privacy-aware recommended tags. The results show that the proposed approach is able to recommend highly relevant tags.

- **Dynamic Deep Multi-modal Fusion for Image Privacy Prediction:** In this chapter, we estimate the competence of object, scene and image tag modalities, derived through convolutional neural networks and dynamically identify the set of most competent modalities for a target image to adequately predict the class of the image as *private* or *public*. The proposed approach contains three stages wherein we first identify neighborhoods for a target image based on visual content similarity and privacy profile similarity. Then, we derive “competence” features from these neighborhoods and provide them to the “competence” classifiers to predict whether a modality is competent for the target image. Lastly, we select the subset of the most competent modalities and take a majority vote to predict privacy class of the target image. Experimental results show that our approach predicts the sensitive (or private) content more accurately than the models trained on an individual modality (object, scene, and tags), multi-modality baselines and prior privacy prediction approaches. Also, our approach could aid other applications such as event understanding, image classification, to on the fly decide which CNN (object, scene or tag) to use based on a target image.

6.2 Summary of Contributions

This section presents the contributions of our works in this dissertation:

1. **The Use of Deep Features for Online Image Sharing.**

- We use three deep feature representations corresponding to the output of three fully-connected layers of an eight-layer deep neural network pre-trained on ILSVRC-2012, a subset of ImageNet dataset consisting of 1.2M+ images labeled with 1,000 object categories¹⁶, as well as the probability distribution over the 1,000 categories obtained from the last layer of the network via softmax.
- As discussed earlier, the set of user tags may be incomplete and noisy. Hence, unlike previous works, we leverage CNNs for automatically generating object tags. We also propose the extraction of scene tags to capture additional information from the visual content that is not captured by existing object tags. We call these object and scene tags as “deep tags.”
- We evaluate the performance of the “deep” features (extracted from AlexNet¹⁷) on a subset of the PicAlert dataset of Flickr images, labeled as private or public. The PicAlert dataset was made publicly available by Zerr et al.⁷.
- We empirically show that learning models trained on deep visual features and deep tags for privacy prediction outperform strong baselines such as those trained on hierarchical deep features, SIFT, GIST (global image descriptors) and user provided tags. We also show that deep visual features provide improved performance for the private class (i.e., correctly identifying more images as private) as compared to baseline approaches.
- Our results show that the deep image tags yield better performing models as compared to user tags and the combination of deep tags and user tags outperforms each set of tags individually.

2. DeepPrivate Features For Image Privacy Prediction.

- We study deep visual semantic features and deep image tags derived from CNN architectures pre-trained on the ImageNet dataset and use them in conjunction with Support Vector Machine (SVM) classifiers for image privacy prediction. Specifically, we extract deep features from four successful (pre-trained) CNN archi-

tures for object recognition, AlexNet, GoogLeNet, VGG-16, and ResNet and compare their performance on the task of privacy prediction. Through carefully designed experiments, we find that ResNet produces the best feature representations for privacy prediction compared with the other CNNs.

- We fine-tune the pre-trained CNN architectures on our privacy dataset and use the softmax function to predict the images' privacy as *public* or *private*. We compare the fine-tuned CNNs with the SVM models obtained on the features derived from the pre-trained CNNs and show that, although the overall performance obtained by the fine-tuned CNNs is comparable to that of SVM models, the fine-tuned networks provide improved recall for the private class as compared to the SVM models trained on the pre-trained features.
- We show that the best feature representation produced by ResNet outperforms several baselines for image privacy prediction that consider CNN-based models and SVM models trained on traditional visual features such as SIFT and global GIST descriptor.
- Next, we investigate the combination of user tags and deep tags derived from CNNs in two settings: (1) using SVM on the bag-of-tags features; and (2) applying the text CNN²¹ on the combination of user tags and deep tags for privacy prediction using the softmax function. We compare these models with the models trained on the most promising visual features extracted from ResNet (obtained from our study) for privacy prediction. Our results show that the models trained on the visual features perform better than those trained on the tag features.
- Finally, we explore the combination of deep visual features with image tags and show further improvement in performance over the individual sets of features.

3. Privacy-Aware Tag Recommendation for Image Sharing.

- We present a privacy-aware approach to automatic image tagging, that aims at improving the quality of user annotations (or user tags), while also preserving the

images' original privacy sharing patterns. Precisely, we recommend potential tags for each target image by mining privacy-aware tags from the most similar images of the target image, which we obtain from a large collection of images.

- We study our privacy-aware recommended tags obtained by the proposed privacy-aware weighting scheme in an ablation experiment for privacy prediction. In this experiment, we compare various privacy-aware and privacy-oblivious weighting schemes and observe how the privacy prediction performance varies for these weighting schemes. We also experiment with various parameter values to estimate the best parameter setting.
- We compare the performance of privacy prediction using tags recommended by the proposed approach against the tags recommended by a prior state-of-the-art image annotation method. Our objective in this experiment is to verify whether the recommended tags by the proposed approach can capture better privacy characteristics than the prior state-of-the-art annotation.
- We investigate tag recommendation in a binary image privacy prediction task and show that the predicted tags can exhibit relevant cues for specific privacy settings (*public* or *private*) that can be used to improve the image privacy prediction performance.
- Our results show that we achieve a better privacy prediction performance when we add the recommended privacy-aware tags to the original user tags and predicted deep tags of images as compared to prior approaches of image privacy prediction.
- We also evaluate the recommended tags by employing crowd-sourcing to identify relevancy of the suggested tags to images. The results show that, although the user-input tags comprise noise or even some images do not have any tags at all, our approach is able to recommend accurate tags. In addition, we evaluate both privacy-aware and privacy-oblivious recommended tags and show that the privacy-aware recommended tags describe an image's content more accurately as compared to the privacy-oblivious tags.

4. Dynamic Deep Multi-modal Fusion for Image Privacy Prediction.

- Our significant contribution is to estimate the competence of object, scene and tag modalities for privacy prediction and dynamically identify the most competent modalities for a target image whose privacy has to be predicted.
- We derive “competence” features from the neighborhood regions of a target image and learn classifiers on them to identify whether a modality is competent to accurately predict the privacy of the target image. To derive these features, we consider privacy and visual neighborhoods of the target image to bring both sensitive and visually similar image content closer.
- We provide an in-depth analysis of our algorithm in an ablation setting, where we record the performance of the proposed approach by removing its various components. The analysis outline the crucial components of our approach.
- Our results show that we identify images’ sensitive content more accurately than single modality models (object, scene, and tag), multi-modality baselines and prior approaches of privacy prediction, depicting that the approach optimally combines the multi-modality for privacy prediction.

6.3 Future Directions

As we mentioned before, the main purpose of image privacy prediction systems is to accurately identify private or sensitive content from images before they are shared on social networking sites. In this study, given an image, we learn models to classify the image into one of the two classes: *private* or *public*, based on generic patterns of privacy. As immediate next steps we plan to:

- Recently, Zhong et al.³² discussed challenges faced by both generic and personalized models for image privacy classification. Specifically, they highlight that generic privacy patterns do not capture well an individual’s sharing behavior, whereas personalized

models generally require large amounts of individual user data to learn reliable models, and are time and space consuming to train and store models for each user. In future, hybrid approaches can be developed that contain both generic and personalized models.

- An architecture can be developed, that will incorporate other contextual information about images such as personal information about the image owner, owner's privacy preferences or the owner social network activities, in addition to the visual content of the image. Another interesting direction is to extend these CNN architectures to describe and localize the sensitive content in private images.
- The tags depicting objects (such as beach, furniture) or actions (such as walking) in images can be identified objectively, whereas abstract tags such as "beautiful," "pretty," etc., are more subjective. These abstract concepts can be studied in personalized privacy settings and can help predict image sensitiveness which explicitly accounts for the variance of human privacy notions.
- It will be interesting to study dynamic multi-modal fusion in personalized privacy settings. Also, other types of competence learning approaches and competence features can be developed for estimating the competence of base classifiers. For example, instead of identifying neighborhood using nearest neighbors, which can be time-consuming, images can be clustered and the most relevant image cluster can be considered for a target image.

Bibliography

- [1] Yann LeCun. Facebook Envisions AI That Keeps You From Uploading Embarrassing Pics. <https://www.wired.com/2014/12/fb/all/1>, 2017. [Online; accessed 12-April-2017].
- [2] Heather Richter Lipford, Andrew Besmer, and Jason Watson. Understanding privacy settings in facebook with an audience view. In *Proceedings of the 1st Conference on Usability, Psychology, and Security*, UPSEC'08, pages 2:1–2:8, 2008.
- [3] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *IEEE International Conference on Computer Vision, ICCV 2017*, pages 3706–3715, 2017.
- [4] Findlaw. Is It Safe to Post Photos of Your Kids Online? <http://consumer.findlaw.com/online-scams/is-it-safe-to-post-photos-of-your-kids-online.html>, 2017. [Online; accessed 3-April-2017].
- [5] Shane Ahern, Dean Eckles, Nathaniel S. Good, Simon King, Mor Naaman, and Rahul Nair. Over-exposed?: Privacy patterns and considerations in online and mobile photo sharing. In *Proceedings of the SIGCHI Conference, CHI '07*, pages 357–366, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-593-9.
- [6] Anna Squicciarini, Smitha Sundareswaran, Dan Lin, and Josh Wede. A3p: adaptive policy prediction for shared images over popular content sharing sites. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, pages 261–270. ACM, 2011.
- [7] Sergej Zerr, Stefan Siersdorfer, Jonathon Hare, and Elena Demidova. Privacy-aware image classification and search. In *Proceedings of the 35th international ACM SIGIR*

- conference on Research and development in information retrieval*, NY, USA, 2012. ACM. ISBN 978-1-4503-1472-5.
- [8] Ralph Gross and Alessandro Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society*, pages 71–80, 2005. ISBN 1-59593-228-3.
- [9] Susan Waters and James Ackerman. Exploring Privacy Management on Facebook: Motivations and Perceived Consequences of Voluntary Disclosure. *Journal of Computer-Mediated Communication*, 17(1):101–115, 2011. ISSN 10836101.
- [10] Mary Madden. Privacy management on social media sites. <http://www.pewinternet.org/2012/02/24/privacy-management-on-social-media-sites>, 2012. [Online; accessed 12-November-2017].
- [11] Sergej Zerr, Stefan Siersdorfer, and Jonathon S. Hare. Picalert!: a system for privacy-aware image classification and retrieval. In Xue wen Chen, Guy Lebanon, Haixun Wang, and Mohammed J. Zaki, editors, *CIKM*, pages 2710–2712. ACM, 2012. ISBN 978-1-4503-1156-4. URL <http://dblp.uni-trier.de/db/conf/cikm/cikm2012.html#ZerrSH12>.
- [12] Anna Squicciarini, Cornelia Caragea, and Rahul Balakavi. Analyzing images’ privacy for the modern web. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT ’14, pages 136–147, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2954-5.
- [13] Anna Squicciarini, Cornelia Caragea, and Rahul Balakavi. Toward automated online photo privacy. *ACM Trans. Web*, 11(1):2:1–2:29, April 2017. ISSN 1559-1131.
- [14] Lam Tran, Deguang Kong, Hongxia Jin, and Ji Liu. Privacy-cnh: A framework to detect photo privacy with convolutional neural network using hierarchical features. In *Proceedings of the Thirtieth AAAI Conference*, pages 1317–1323, 2016.

- [15] Elena Zheleva and Lise Getoor. To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th International Conference on World Wide Web, WWW '09*, pages 531–540, 2009.
- [16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, pages 1–42, April 2015.
- [17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
- [18] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.
- [19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL <https://doi.org/10.1109/CVPR.2016.90>.
- [21] Yoon Kim. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751. ACL, 2014.
- [22] Ashwini Tonge. Identifying private content for online image sharing. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI*

- Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 8040–8041, 2018. URL <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17065>.
- [23] H. Sundaram, L. Xie, M. De Choudhury, Y.R. Lin, and A. Natsev. Multimedia semantics: Interactions between content and community. *Proceedings of the IEEE*, 100(9):2737–2758, 2012.
- [24] Ashwini Tonge and Cornelia Caragea. Image privacy prediction using deep features. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 4266–4267, 2016.
- [25] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*, 2016.
- [26] Ashwini Tonge, Cornelia Caragea, and Anna Squicciarini. Uncovering scene context for predicting privacy of online shared images. In *AAAI’ 18*, 2018.
- [27] Ashwini Tonge and Cornelia Caragea. On the use of “deep” features for online image sharing. In *The Web Conference Companion*, 2018.
- [28] Ashwini Tonge, Cornelia Caragea, and Anna Squicciarini. Privacy-aware tag recommendation for image sharing. In *Proceedings of the 29th on Hypertext and Social Media, HT ’18*, pages 52–56, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5427-1. doi: 10.1145/3209542.3209574. URL <http://doi.acm.org/10.1145/3209542.3209574>.
- [29] Ashwini Tonge and Cornelia Caragea. Dynamically identifying deep multimodal features for image privacy prediction. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 2019.
- [30] Ashwini Tonge and Cornelia Caragea. Dynamic deep multi-modal fusion for image privacy prediction. In *The Web Conference*, 2019.

- [31] Benjamin Laxton, Kai Wang, and Stefan Savage. Reconsidering physical key secrecy: Teleduplication via optical decoding. In *Proceedings of the 15th ACM Conference on Computer and Communications Security, CCS '08*, pages 469–478. ACM, 2008. ISBN 978-1-59593-810-7.
- [32] Haoti Zhong, Anna Squicciarini, David Miller, and Cornelia Caragea. A group-based personalized model for image privacy classification and labeling. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3952–3958, 2017.
- [33] Daniel Buschek, Moritz Bader, Emanuel von Zezschwitz, and Alexander De Luca. Automatic privacy classification of personal photos. In Julio Abascal, Simone Barbosa, Mirko Fetter, Tom Gross, Philippe Palanque, and Marco Winckler, editors, *Human Computer Interaction INTERACT 2015*, volume 9297, pages 428–435. 2015. ISBN 978-3-319-22667-5.
- [34] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, November 2004. ISSN 0920-5691.
- [35] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, May 2001. ISSN 0920-5691.
- [36] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR 2014*. CBLS, April 2014.
- [37] Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala, and Yann Lecun. Pedestrian detection with unsupervised multi-stage feature learning. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3626–3633, 2013. ISBN 978-0-7695-4989-7.
- [38] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning

- hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, August 2013.
- [39] Sergey Karayev, Aaron Hertzmann, Holger Winnemoeller, Aseem Agarwala, and Trevor Darrell. Recognizing image style. *CoRR*, abs/1311.3715, 2013.
- [40] Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, Adrian Popescu, and Yiannis Kompatsiaris. Personalized privacy-aware image classification. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ICMR '16*, pages 71–78, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4359-6.
- [41] Jun Yu, Baopeng Zhang, Zhengzhong Kuang, Dan Lin, and Jianping Fan. iprivacy: Image privacy protection by identifying sensitive objects via deep multi-task learning. *IEEE Trans. Information Forensics and Security*, 12(5):1005–1016, 2017.
- [42] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678, 2014. ISBN 978-1-4503-3063-3.
- [43] Paul Viola and Michael Jones. Robust real-time object detection. In *IJCV*, 2001.
- [44] Haitao Xu, Haining Wang, and Angelos Stavrou. Privacy risk assessment on online photos. In Herbert Bos, Fabian Monrose, and Gregory Blanc, editors, *RAID*, volume 9404, pages 427–447, 2015. ISBN 978-3-319-26361-8.
- [45] Benjamin Henne, Christian Szongott, and Matthew Smith. Snapme if you can: Privacy threats of other peoples’ geo-tagged media and what we can do about it. *WiSec '13*, 2013. ISBN 978-1-4503-1998-0. doi: 10.1145/2462096.2462113.
- [46] Balachander Krishnamurthy and Craig E. Wills. Characterizing privacy in online social networks. In *Proceedings of the First Workshop on Online Social Networks, WOSN '08*, pages 37–42, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-182-8. doi: 10.1145/1397735.1397744. URL <http://doi.acm.org/10.1145/1397735.1397744>.

- [47] Andrew Simpson. On the need for user-defined fine-grained access control policies for social networking applications. In *Proceedings of the Workshop on Security in Opportunistic and SOcial Networks*, SOSOC '08, pages 1:1–1:8, New York, NY, USA, 2008. ACM. doi: 10.1145/1461469.1461470. URL <http://doi.acm.org/10.1145/1461469.1461470>.
- [48] Kambiz Ghazinour, Stan Matwin, and Marina Sokolova. Monitoring and recommending privacy settings in social networks. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, EDBT '13, pages 164–168, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1599-9. doi: 10.1145/2457317.2457344. URL <http://doi.acm.org/10.1145/2457317.2457344>.
- [49] Javier Parra-Arnau, David Rebollo-Monedero, Jordi Forné, Jose L. Muñoz, and Oscar Esparza. Optimal tag suppression for privacy protection in the semantic web. *Data Knowl. Eng.*, 81-82:46–66, 2012.
- [50] J. Parra-Arnau, A. Perego, E. Ferrari, J. Forn, and D. Rebollo-Monedero. Privacy-preserving enhanced collaborative tagging. *IEEE Transactions on Knowledge and Data Engineering*, 26(1):180–193, Jan 2014. ISSN 1041-4347. doi: 10.1109/TKDE.2012.248.
- [51] Panagiotis Ilia, Iasonas Polakis, Elias Athanasopoulos, Federico Maggi, and Sotiris Ioannidis. Face/off: Preventing privacy leakage from photos in social networks. In *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pages 781–792, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3832-5. doi: 10.1145/2810103.2813603. URL <http://doi.acm.org/10.1145/2810103.2813603>.
- [52] Simon Jones and Eamonn O'Neill. Contextual dynamics of group-based sharing decisions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1777–1786. ACM, 2011. ISBN 978-1-4503-0228-9.
- [53] Andrew Besmer and Heather Lipford. Tagged photos: concerns, perceptions, and

- protections. In *CHI '09: 27th international conference extended abstracts on Human factors in computing systems*, pages 4585–4590, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-247-4.
- [54] João Paulo Pesce, Diego Las Casas, Gustavo Rauber, and Virgílio Almeida. Privacy attacks in social media using photo tagging networks: A case study with facebook. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media, PSOSM '12*, pages 4:1–4:8, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1236-3.
- [55] Delphine Christin, Pablo SáNchez López, Andreas Reinhardt, Matthias Hollick, and Michaela Kauer. Share with strangers: Privacy bubbles as user-centered privacy control for mobile content sharing applications. *Inf. Secur. Tech. Rep.*, 17(3):105–116, February 2013. ISSN 1363-4127.
- [56] Mohammad Mannan and Paul C. van Oorschot. Privacy-enhanced sharing of personal content on the web. In *Proceedings of the 17th International Conference on World Wide Web*, pages 487–496, 2008. ISBN 978-1-60558-085-2.
- [57] James B. D. Joshi and Tao Zhang, editors. *The 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing, CollaborateCom 2009, Washington DC, USA, November 11-14, 2009*, 2009. ICST / IEEE. ISBN 978-963-9799-76-9.
- [58] Anna Squicciarini, Mohamed Shehab, and Federica Paci. Collective privacy management in social networks. In *Proceedings of the 18th International Conference on World Wide Web, WWW '09*, pages 521–530, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-487-4.
- [59] Joseph Bonneau, Jonathan Anderson, and Luke Church. Privacy suites: Shared privacy for social networks. In *Proceedings of the 5th Symposium on Usable Privacy and Security, SOUPS '09*, pages 30:1–30:1, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-736-3.

- [60] Joseph Bonneau, Jonathan Anderson, and George Danezis. Prying data out of a social network. In *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining, ASONAM '09*, pages 249–254, Washington, DC, USA, 2009. IEEE Computer Society. ISBN 978-0-7695-3689-7.
- [61] Lujun Fang and Kristen LeFevre. Privacy wizards for social networking sites. In *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, pages 351–360, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8.
- [62] Anna Squicciarini, D. Lin, S. Karumanchi, and N. DeSisto. Automatic social group organization and privacy management. In *8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*, pages 89–96, Oct 2012.
- [63] George Danezis. Inferring privacy policies for social networking services. In *Proceedings of the 2nd ACM Workshop on Security and Artificial Intelligence, AISec '09*, pages 5–10, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-781-3. doi: 10.1145/1654988.1654991. URL <http://doi.acm.org/10.1145/1654988.1654991>.
- [64] Jason Watson, Heather Richter Lipford, and Andrew Besmer. Mapping user preference to privacy default settings. *ACM Trans. Comput.-Hum. Interact.*, 22(6):32:1–32:20, November 2015. ISSN 1073-0516. doi: 10.1145/2811257. URL <http://doi.acm.org/10.1145/2811257>.
- [65] Berkant Kepez and Pinar Yolum. Learning privacy rules cooperatively in online social networks. In *Proceedings of the 1st International Workshop on AI for Privacy and Security, PrAISe '16*, pages 3:1–3:4, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4304-6. doi: 10.1145/2970030.2970036. URL <http://doi.acm.org/10.1145/2970030.2970036>.
- [66] Georgios Petkos, Symeon Papadopoulos, and Yiannis Kompatsiaris. Social circle discovery in ego-networks by mining the latent structure of user connections and profile

- attributes. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pages 880–887, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3854-7.
- [67] Anna Squicciarini, Dan Lin, Smitha Sundareswaran, and Joshua Wede. Privacy policy inference of user-uploaded images on content sharing sites. *IEEE Trans. Knowl. Data Eng.*, 27(1):193–206, 2015.
- [68] Peter F. Klemperer, Yuan Liang, Michelle L. Mazurek, Manya Sleeper, Blase Ur, Lujo Bauer, Lorrie Faith Cranor, Nitin Gupta, and Michael K. Reiter. Tag, you can see it! Using tags for access control in photo sharing. In *CHI 2012: Conference on Human Factors in Computing Systems*. ACM, May 2012.
- [69] Lin Yuan, Joel Regis Theytaz, and Touradj Ebrahimi. Context-dependent privacy-aware photo sharing based on machine learning. *Proc. of 32nd International Conference on ICT Systems Security and Privacy Protection (IFIP SEC 2017)*, 2017.
- [70] Fabeah Adu-Oppong, Casey K. Gardiner, Apu Kapadia, and Patrick P. Tsang. Social circles: Tackling privacy in social networks. In *Symposium on Usable Privacy and Security (SOUPS)*, 2008.
- [71] Katarzyna Olejnik, Italo Dacosta, Joana Soares Machado, Kévin Huguenin, Mohammad Emtiyaz Khan, and Jean-Pierre Hubaux. SmarPer: Context-Aware and Automatic Runtime-Permissions for Mobile Devices. In *38th IEEE Symposium on Security and Privacy (S&P)*, pages 1058–1076, San Jose, CA, United States, May 2017. IEEE. doi: 10.1109/SP.2017.25. URL <https://hal.archives-ouvertes.fr/hal-01489684>.
- [72] Igor Bilogrevic, Kévin Huguenin, Berker Agir, Murtuza Jadliwala, Maria Gazaki, and Jean-Pierre Hubaux. A machine-learning based approach to privacy-aware information-sharing in mobile social networks. *Pervasive and Mobile Computing*, 25:125–142, 2016.
- [73] Julien Freudiger, Reza Shokri, and Jean-Pierre Hubaux. Evaluating the privacy risk of location-based services. In George Danezis, editor, *Financial Cryptography and Data*

- Security*, pages 31–46, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-27576-0.
- [74] Arjun Baokar. A contextually-aware, privacy-preserving android permission model. In *Technical Report No. UCB/EECS-2016-69, University of California, Berkeley*, 2016.
- [75] Drew Fisher, Leah Dorner, and David Wagner. Short paper: Location privacy: User behavior in the field. In *Proceedings of the Second ACM Workshop on Security and Privacy in Smartphones and Mobile Devices, SPSM '12*, pages 51–56, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1666-8. doi: 10.1145/2381934.2381945. URL <http://doi.acm.org/10.1145/2381934.2381945>.
- [76] Gerald Friedland and Robin Sommer. Cybercasing the joint: On the privacy implications of geo-tagging. In *HotSec*. USENIX Association, 2010.
- [77] Ramprasad Ravichandran, Michael Benisch, Patrick Gage Kelley, and Norman M. Sadeh. *Capturing Social Networking Privacy Preferences*., pages 1–18. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-642-03168-7.
- [78] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. Quantifying location privacy. In *IEEE Symposium on Security and Privacy*, pages 247–262. IEEE Computer Society, 2011.
- [79] Eran Toch. Crowdsourcing privacy preferences in context-aware applications. *Personal Ubiquitous Comput.*, 18(1):129–141, January 2014. ISSN 1617-4909. doi: 10.1007/s00779-012-0632-0. URL <http://dx.doi.org/10.1007/s00779-012-0632-0>.
- [80] Yuchen Zhao, Juan Ye, and Tristan Henderson. Privacy-aware location privacy preference recommendations. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, MOBIQUITOUS '14*, pages 120–129, ICST, Brussels, Belgium, Belgium, 2014. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineer-

- ing). ISBN 978-1-63190-039-6. doi: 10.4108/icst.mobiquitous.2014.258017. URL <http://dx.doi.org/10.4108/icst.mobiquitous.2014.258017>.
- [81] Jaeyoung Choi, Martha Larson, Xinchao Li, Kevin Li, Gerald Friedland, and Alan Hanjalic. The geo-privacy bonus of popular photo enhancements. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR '17*, pages 84–92, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4701-3. doi: 10.1145/3078971.3080543. URL <http://doi.acm.org/10.1145/3078971.3080543>.
- [82] C.M.A. Yeung, L. Kagal, N. Gibbins, and N. Shadbolt. Providing access control to online photo albums based on tags and linked data. *Social Semantic Web: Where Web*, 2, 2009.
- [83] Emilia Apostolova and Dina Demner-Fushman. Towards automatic image region annotation - image region textual coreference resolution. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 41–44. Association for Computational Linguistics, 2009.
- [84] Nitya Vyas, Anna Squicciarini, Chih-Cheng Chang, and Danfeng Yao. Towards automatic privacy management in web 2.0 with semantic analysis on annotations. In *CollaborateCom*, pages 1–10, 2009.
- [85] Munmun De Choudhury, Hari Sundaram, Yu-Ru Lin, Ajita John, and Doree Duncan Seligmann. Connecting content to community in social media via image content, user tags and user communication. In *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo, ICME'09*, pages 1238–1241, Piscataway, NJ, USA, 2009. IEEE Press. ISBN 978-1-4244-4290-4. URL <http://dl.acm.org/citation.cfm?id=1698924.1699229>.
- [86] Moo-Ryong Ra, Ramesh Govindan, and Antonio Ortega. P3: Toward privacy-preserving photo sharing. In *Proceedings of the 10th USENIX Conference on Networked*

- Systems Design and Implementation*, nsdi'13, pages 515–528, Berkeley, CA, USA, 2013. USENIX Association. URL <http://dl.acm.org/citation.cfm?id=2482626.2482675>.
- [87] Abdurrahman Can Kurtan and Pinar Yolum. Pelte: Privacy estimation of images from tags. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pages 1989–1991, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems. URL <http://dl.acm.org/citation.cfm?id=3237383.3238047>.
- [88] Anna Squicciarini, Andrea Novelli, Dan Lin, Cornelia Caragea, and Haoti Zhong. From tag to protect: A tag-driven policyrecommender system for image sharing. In *PST '17*, 2017.
- [89] Zhenzhong Kuang, Zongmin Li, Dan Lin, and Jianping Fan. Automatic privacy prediction to accelerate social image sharing. In *Third IEEE International Conference on Multimedia Big Data, BigMM 2017, Laguna Hills, CA, USA, April 19-21, 2017*, pages 197–200, 2017. doi: 10.1109/BigMM.2017.70. URL <https://doi.org/10.1109/BigMM.2017.70>.
- [90] Jun Yu, Zhenzhong Kuang, Zhou Yu, Dan Lin, and Jianping Fan. Privacy setting recommendation for image sharing. In *16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017, Cancun, Mexico, December 18-21, 2017*, pages 726–730, 2017. doi: 10.1109/ICMLA.2017.00-73. URL <https://doi.org/10.1109/ICMLA.2017.00-73>.
- [91] Jun Yu, Zhenzhong Kuang, Baopeng Zhang, Wei Zhang, Dan Lin, and Jianping Fan. Leveraging content sensitiveness and user trustworthiness to recommend fine-grained privacy settings for social image sharing. *IEEE Trans. Information Forensics and Security*, 13(5):1317–1332, 2018. doi: 10.1109/TIFS.2017.2787986. URL <https://doi.org/10.1109/TIFS.2017.2787986>.

- [92] Wei Zhang, S. S. Cheung, and Minghua Chen. Hiding privacy information in video surveillance system. In *IEEE International Conference on Image Processing 2005*, volume 3, pages II–868, Sept 2005. doi: 10.1109/ICIP.2005.1530530.
- [93] Yuta Nakashima, Noboru Babaguchi, and Jianping Fan. Privacy protection for social video via background estimation and crf-based videographer’s intention modeling. *IEICE Transactions*, 99-D(4):1221–1233, 2016.
- [94] Yuta Nakashima, Noboru Babaguchi, and Jianping Fan. Intended human object detection for automatically protecting privacy in mobile video surveillance. *Multimedia Syst.*, 18(2):157–173, 2012. doi: 10.1007/s00530-011-0244-y. URL <https://doi.org/10.1007/s00530-011-0244-y>.
- [95] Yuta Nakashima, Noboru Babaguchi, and Jianping Fan. Automatic generation of privacy-protected videos using background estimation. In *Proceedings of the 2011 IEEE International Conference on Multimedia and Expo, ICME 2011, 11-15 July, 2011, Barcelona, Catalonia, Spain*, pages 1–6, 2011. doi: 10.1109/ICME.2011.6011955. URL <https://doi.org/10.1109/ICME.2011.6011955>.
- [96] F. Dufaux and T. Ebrahimi. Scrambling for privacy protection in video surveillance systems. *IEEE Trans. Cir. and Sys. for Video Technol.*, 18(8):1168–1174, August 2008. ISSN 1051-8215. doi: 10.1109/TCSVT.2008.928225. URL <https://doi.org/10.1109/TCSVT.2008.928225>.
- [97] Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [98] D. Hu, F. Chen, X. Wu, and Z. Zhao. A framework of privacy decision recommendation for image sharing in online social networks. In *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)*, pages 243–251, June 2016.

- [99] S. Shamma and M. Y. S. Uddin. Towards privacy-aware photo sharing using mobile phones. In *8th International Conference on Electrical and Computer Engineering*, pages 449–452, Dec 2014. doi: 10.1109/ICECE.2014.7026919.
- [100] X. Yuan, X. Wang, C. Wang, Anna Squicciarini, and K. Ren. Towards privacy-preserving and practical image-centric social discovery. *IEEE Transactions on Dependable and Secure Computing*, 15(5):868–882, Sept 2018. ISSN 1545-5971.
- [101] Emanuel von Zezschwitz, Sigrid Ebbinghaus, Heinrich Hussmann, and Alexander De Luca. You can’t watch this!: Privacy-respectful photo browsing on smartphones. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI ’16, pages 4320–4324, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858120. URL <http://doi.acm.org/10.1145/2858036.2858120>.
- [102] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, volume 11220 of *Lecture Notes in Computer Science*, pages 627–645. Springer, 2018. doi: 10.1007/978-3-030-01270-0_37. URL https://doi.org/10.1007/978-3-030-01270-0_37.
- [103] Erika McCallister, Timothy Grance, and Karen A. Scarfone. Sp 800-122. guide to protecting the confidentiality of personally identifiable information (pii). Technical report, Gaithersburg, MD, United States, 2010.
- [104] Yifang Li, Wyatt Troutman, Bart P. Knijnenburg, and Kelly Caine. Human perceptions of sensitive content in photos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.
- [105] Haoti Zhong, Anna Squicciarini, and David Miller. Toward automated multiparty

- privacy conflict detection. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pages 1811–1814, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-6014-2.
- [106] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016.
- [107] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *CoRR*, abs/1310.1531, 2013.
- [108] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham W. Taylor, and Daniel L. Silver, editors, *ICML Unsupervised and Transfer Learning*, volume 27 of *JMLR Proceedings*, pages 17–36. JMLR.org, 2012. URL <http://dblp.uni-trier.de/db/journals/jmlr/jmlrp27.html#Bengio12>.
- [109] Kerstin Bischoff, Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu. Can All Tags Be Used for Search? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 193–202, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458112. URL <http://dx.doi.org/10.1145/1458082.1458112>.
- [110] Livia Hollenstein and Ross Purves. Exploring place through user-generated content: Using flickr tags to describe city cores. *J. Spatial Information Science*, 1(1):21–48, 2010. doi: 10.5311/JOSIS.2010.1.3. URL <https://doi.org/10.5311/JOSIS.2010.1.3>.
- [111] Jinhui Tang, Shuicheng Yan, Richang Hong, Guo-Jun Qi, and Tat-Seng Chua. Inferring semantic concepts from community-contributed images and noisy tags. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, pages 223–232, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-608-3. doi: 10.1145/1631272.1631305. URL <http://doi.acm.org/10.1145/1631272.1631305>.

- [112] Yue Gao, Meng Wang, Huanbo Luan, Jialie Shen, Shuicheng Yan, and Dacheng Tao. Tag-based social image search with visual-text joint hypergraph learning. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, pages 1517–1520, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0616-4. doi: 10.1145/2072298.2072054. URL <http://doi.acm.org/10.1145/2072298.2072054>.
- [113] Wen-tau Yih, Xiaodong He, and Christopher Meek. Semantic parsing for single-relation question answering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 643–648, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/P14-2105>.
- [114] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *CoRR*, abs/1404.2188, 2014.
- [115] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November 2011. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=1953048.2078186>.
- [116] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1188–1196. JMLR.org, 2014.
- [117] Bullguard. Privacy violations, the dark side of social media, 2018. <http://www.bullguard.com/bullguard-security-center/internet-security/social-media-dangers/privacy-violations-in-social-media.aspx>.
- [118] Shaolei Feng, Raghavan Manmatha, and Victor Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR (2)*, pages 1002–1009, 2004.
- [119] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. Tag-

- prop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009.
- [120] Jing Liu, Mingjing Li, Qingshan Liu, Hanqing Lu, and Songde Ma. Image annotation via graph learning. *PR*, February 2009. ISSN 0031-3203.
- [121] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. A new baseline for image annotation. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Computer Vision – ECCV 2008*, pages 316–329. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-88690-7.
- [122] Minmin Chen, Alice Zheng, and Kilian Q. Weinberger. Fast image tagging. In *Proceedings of the 30th International Conference on Machine Learning. ICML*, January 2013. URL <https://www.microsoft.com/en-us/research/publication/fast-image-tagging/>.
- [123] Alexei Yavlinsky, Edward Schofield, and Stefan Ruger. *Automated Image Annotation Using Global Features and Robust Nonparametric Density Estimation*. Springer Berlin Heidelberg, 2005. ISBN 978-3-540-31678-7.
- [124] Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. in AI*, 2009. ISSN 1687-7470.
- [125] Victor Lavrenko, R. Manmatha, and Jiwoon Jeon. A model for learning the semantics of pictures. In *NIPS 16*. MIT Press, 2004.
- [126] Chong Wang, David M. Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *CVPR 2009*, pages 1903–1910. IEEE, 2009.
- [127] David M. Blei and Michael I. Jordan. Modeling annotated data. *SIGIR ’03*, pages 127–134, New York, NY, USA, 2003. ACM. ISBN 1-58113-646-3.
- [128] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135,

- March 2003. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=944919.944965>.
- [129] Yansong Feng and Mirella Lapata. Automatic image annotation using auxiliary text information. In *ACL-08: HLT*, Columbus, Ohio, June 2008.
- [130] Yansong Feng and Mirella Lapata. Topic models for image annotation and text illustration. *HLT '10*, pages 831–839, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5.
- [131] P. Duygulu, Kobus Barnard, J. F. G. de Freitas, and David A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision-Part IV, ECCV '02*, pages 97–112, London, UK, UK, 2002. Springer-Verlag. ISBN 3-540-43748-7. URL <http://dl.acm.org/citation.cfm?id=645318.649254>.
- [132] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, SIGIR '03*, pages 119–126, New York, NY, USA, 2003. ACM. ISBN 1-58113-646-3.
- [133] Florent Monay and Daniel Gatica-Perez. Plsa-based image auto-annotation: Constraining the latent space. In *Proceedings of the 12th Annual ACM International Conference on Multimedia, MULTIMEDIA '04*, pages 348–351. ACM, 2004. ISBN 1-58113-893-8.
- [134] Yuxin Peng, Zhiwu Lu, and Jianguo Xiao. Semantic concept annotation based on audio plsa model. In *Proceedings of the 17th ACM International Conference on Multimedia, MM '09*, pages 841–844, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-608-3.
- [135] Duangmanee Putthividhya, Hagai Attias, and Srikantan S. Nagarajan. Topic regression multi-modal latent dirichlet allocation for image annotation. *2010 IEEE Computer*

- Society Conference on Computer Vision and Pattern Recognition*, pages 3408–3415, 2010.
- [136] Rainer Lienhart, Stefan Romberg, and Eva Hörster. Multilayer plsa for multimodal image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '09, pages 9:1–9:8, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-480-5.
- [137] Arnab Ghoshal, Pavel Ircing, and Sanjeev Khudanpur. Hidden markov models for automatic annotation and content-based retrieval of images and video. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 544–551, New York, NY, USA, 2005. ACM. ISBN 1-59593-034-5.
- [138] F. Yu and H. H. S Ip. Automatic semantic annotation of images using spatial hidden markov model. In *2006 IEEE International Conference on Multimedia and Expo*, pages 305–308, July 2006.
- [139] Yufeng Zhao, Yao Zhao, and Zhenfeng Zhu. Tsvm-hmm: Transductive svm based hidden markov model for automatic image annotation. *Expert Systems with Applications*, 36(6):9813 – 9818, 2009. ISSN 0957-4174.
- [140] Gianluigi Ciocca, Claudio Cusano, Simone Santini, and Raimondo Schettini. Halfway through the semantic gap: Prosemantic features for image retrieval. *Inf. Sci.*, 181(22):4943–4958, November 2011. ISSN 0020-0255.
- [141] E. Chang, Kingshy Goh, G. Sychay, and Gang Wu. Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(1):26–38, Jan 2003. ISSN 1051-8215.
- [142] D. Grangier and S. Bengio. A discriminative kernel-based approach to rank images from text queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1371–1384, Aug 2008. ISSN 0162-8828.

- [143] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):394–410, March 2007. ISSN 0162-8828.
- [144] Venkatesh N. Murthy, Ethem F. Can, and R. Manmatha. A hybrid model for automatic image annotation. In *Proceedings of International Conference on Multimedia Retrieval, ICMR '14*, pages 369:369–369:376, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2782-4.
- [145] Ivica Dimitrovski, Dragi Kocev, Suzana Loskovska, and Sao Deroski. Hierarchical annotation of medical images. *Pattern Recognition*, 44(10):2436 – 2449, 2011. ISSN 0031-3203. Semi-Supervised Learning for Visual Content Analysis and Understanding.
- [146] Hua Wang, Heng Huang, and C. Ding. Image annotation using multi-label correlated green’s function. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2029–2034, Sept 2009.
- [147] H. Wang and J. Hu. Multi-label image annotation via maximum consistency. In *2010 IEEE International Conference on Image Processing*, pages 2337–2340, Sept 2010.
- [148] H. Wang, H. Huang, and C. Ding. Image annotation using bi-relational graph of images and semantic labels. In *CVPR 2011*, pages 793–800, June 2011.
- [149] Gang Chen, Yangqiu Song, Fei Wang, and Changshui Zhang. *Semi-supervised Multi-label Learning by Solving a Sylvester Equation*, pages 410–419. 2008.
- [150] L. Feng and B. Bhanu. Semantic concept co-occurrence patterns for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):785–799, April 2016. ISSN 0162-8828.
- [151] Zheng-Jun Zha, Tao Mei, Jingdong Wang, Zengfu Wang, and Xian-Sheng Hua. Graph-based semi-supervised learning with multiple labels. *Journal of Visual Communication and Image Representation*, 20(2):97 – 103, 2009. ISSN 1047-3203. Special issue on Emerging Techniques for Multimedia Content Sharing, Search and Understanding.

- [152] B. Bao, T. Li, and S. Yan. Hidden-concept driven multilabel image annotation and label ranking. *IEEE Transactions on Multimedia*, 14(1):199–210, Feb 2012. ISSN 1520-9210.
- [153] Sheng-Jun Huang and Zhi-Hua Zhou. Multi-label learning by exploiting label correlations locally. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI’12, pages 949–955. AAAI Press, 2012.
- [154] X. Jing, F. Wu, Z. Li, R. Hu, and D. Zhang. Multi-label dictionary learning for image annotation. *IEEE Transactions on Image Processing*, 25(6):2712–2725, June 2016. ISSN 1057-7149.
- [155] Baoyuan Wu, Siwei Lyu, Bao-Gang Hu, and Qiang Ji. Multi-label learning with missing labels for image annotation and facial action unit recognition. *Pattern Recognition*, 48(7):2279 – 2289, 2015. ISSN 0031-3203.
- [156] Zechao Li, Jing Liu, Changsheng Xu, and Hanqing Lu. Mrank: Multi-correlation learning to rank for image annotation. *Pattern Recognition*, 46(10):2700 – 2710, 2013. ISSN 0031-3203.
- [157] Marina Ivasic-Kos, Miran Pobar, and Slobodan Ribaric. Two-tier image annotation model based on a multi-label classifier and fuzzy-knowledge representation scheme. *Pattern Recognition*, 52:287 – 305, 2016. ISSN 0031-3203.
- [158] Lei Wu, Rong Jin, and Anil K Jain. Tag completion for image retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 35(3), March 2013. ISSN 0162-8828.
- [159] Z. Qin, C. Li, H. Zhang, and J. Guo. Improving tag matrix completion for image annotation and retrieval. In *2015 Visual Communications and Image Processing (VCIP)*, pages 1–4, Dec 2015.
- [160] Z. Lin, G. Ding, M. Hu, J. Wang, and X. Ye. Image tag completion via image-specific and tag-specific linear sparse reconstructions. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1618–1625, June 2013.

- [161] Zijia Lin, Guiguang Ding, Mingqing Hu, Yunzhen Lin, and Shuzhi Sam Ge. Image tag completion via dual-view linear sparse reconstructions. *Computer Vision and Image Understanding*, 124:42 – 60, 2014. ISSN 1077-3142.
- [162] Y. Hou and X. Zhang. A geometric constrained hcrf for object recognition. In *2015 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, pages 1–4, Sept 2015.
- [163] X. Li, Y. Zhang, B. Shen, and B. Liu. Image tag completion by low-rank factorization with dual reconstruction structure preserved. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 3062–3066, Oct 2014.
- [164] X. Li, B. Shen, B. Liu, and Y. Zhang. A locality sensitive low-rank model for image tag completion. *IEEE Transactions on Multimedia*, 18(3):474–483, March 2016. ISSN 1520-9210.
- [165] H. Hu, G. Zhou, Z. Deng, Z. Liao, and G. Mori. Learning structured inference neural networks with label relations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2960–2968, June 2016.
- [166] Y. Niu, Z. Lu, J. Wen, T. Xiang, and S. Chang. Multi-modal multi-scale deep learning for large-scale image annotation. *IEEE Transactions on Image Processing*, pages 1–1, 2018. ISSN 1057-7149.
- [167] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *CoRR*, abs/1312.4894, 2013. URL <http://arxiv.org/abs/1312.4894>.
- [168] Ronggui Wang, Yunfei Xie, Juan Yang, Lixia Xue, Min Hu, and Qingyang Zhang. Large scale automatic image annotation based on convolutional neural network. *Journal of Visual Communication and Image Representation*, 49:213 – 224, 2017. ISSN 1047-3203.

- [169] Yang Yang, Wensheng Zhang, and Yuan Xie. Image automatic annotation via multi-view deep representation. *Journal of Visual Communication and Image Representation*, 33:368 – 377, 2015. ISSN 1047-3203.
- [170] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu. Cnn-rnn: A unified framework for multi-label image classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2285–2294, June 2016.
- [171] Jiren Jin and H. Nakayama. Annotation order matters: Recurrent image annotator for arbitrary length image tagging. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2452–2457, Dec 2016.
- [172] J. Wu, Yinan Yu, Chang Huang, and Kai Yu. Deep multiple instance learning for image classification and auto-annotation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3460–3469, June 2015.
- [173] Lei Wu, Steven C.H. Hoi, Rong Jin, Jianke Zhu, and Nenghai Yu. Distance metric learning from uncertain side information with application to automated photo tagging. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, pages 135–144, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-608-3.
- [174] Pengcheng Wu, Steven Chu-Hong Hoi, Peilin Zhao, and Ying He. Mining social images with distance metric learning for automated image tagging. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 197–206, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0493-1.
- [175] P. Bakliwal and C. V. Jawahar. Active learning based image annotation. In *2015 Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, pages 1–4, Dec 2015.
- [176] F. Tian and X. Shen. Learning label set relevance for search based image annotation. In *2014 International Conference on Virtual Reality and Visualization*, pages 260–265, Aug 2014.

- [177] Zijia Lin, Guiguang Ding, Mingqing Hu, Jianmin Wang, and Jianguang Sun. Automatic image annotation using tag-related random search over visual neighbors. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12*, pages 1784–1788, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1156-4.
- [178] M. M. Kalayeh, H. Idrees, and M. Shah. Nmf-knn: Image annotation using weighted multi-view non-negative matrix factorization. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 184–191, June 2014.
- [179] Qimin Cheng, Qian Zhang, Peng Fu, Conghuan Tu, and Sen Li. A survey and analysis on automatic image annotation. *Pattern Recognition*, 79:242 – 259, 2018. ISSN 0031-3203.
- [180] Zhichen Xu, Yun Fu, Jianchang Mao, and Difu Su. Towards the semantic web: Collaborative tag suggestions. In *Collaborative Web Tagging Workshop at 15th Int. WWW Conference*, 2006.
- [181] Börkur Sigurbjörnsson and Roelof van Zwol. Flickr tag recommendation based on collective knowledge. *WWW '08*, pages 327–336, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2.
- [182] Jing Peng, Daniel Dajun Zeng, Huimin Zhao, and Fei-yue Wang. Collaborative filtering in social tagging systems based on joint item-tag recommendations. *CIKM '10*, pages 809–818. ACM, 2010. ISBN 978-1-4503-0099-5.
- [183] Paul Seitlinger, Dominik Kowald, Christoph Trattner, and Tobias Ley. Recommending tags with a model of human categorization. *CIKM '13*, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2263-8.
- [184] J. Liu, Z. Li, J. Tang, Y. Jiang, and H. Lu. Personalized geo-specific tag recommendation for photos on social websites. *IEEE Transactions on Multimedia*, 16(3):588–600, April 2014. ISSN 1520-9210.

- [185] G. Toderici, H. Aradhye, M. Pasca, L. Sbaiz, and J. Yagnik. Finding meaning on youtube: Tag recommendation and category discovery. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3447–3454, June 2010.
- [186] Yuyun Gong and Qi Zhang. Hashtag recommendation using attention-based convolutional neural network. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI’16*, pages 2782–2788. AAAI Press, 2016. ISBN 978-1-57735-770-4.
- [187] Qi Zhang, Jiawen Wang, Haoran Huang, Xuanjing Huang, and Yeyun Gong. Hashtag recommendation for multimodal microblog using co-attention network. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3420–3426, 2017.
- [188] Hanh T. H. Nguyen, Martin Wistuba, and Lars Schmidt-Thieme. Personalized tag recommendation for images using deep transfer learning. In *Machine Learning and Knowledge Discovery in Databases*, pages 705–720, Cham, 2017. Springer International Publishing. ISBN 978-3-319-71246-8.
- [189] Boon-Siew Seah, Aixin Sun, and Sourav S. Bhowmick. Killing two birds with one stone: Concurrent ranking of tags and comments of social images. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 937–940, 2018.
- [190] Xuemeng Song, Xiang Wang, Liqiang Nie, Xiangnan He, Zhumin Chen, and Wei Liu. A personal privacy preserving framework: I let you know who can see what. In *SIGIR*, pages 295–304. ACM, 2018.
- [191] D. Keerthi Chandra, W. Chowgule, Y. Fu, and D. Lin. Ripa: Real-time image privacy alert system. In *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*, pages 136–145, Oct 2018.

- [192] Yue Shi, Martha Larson, and Alan Hanjalic. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Comput. Surv.*, May 2014. ISSN 0360-0300.
- [193] Christiane Fellbaum. ed. wordnet: an electronic lexical database. *MIT Press, Cambridge MA*, 1998.
- [194] M. B. Mayhew, B. Chen, and K. S. Ni. Assessing semantic information in convolutional neural network representations of images via image annotation. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2266–2270, Sep. 2016.
- [195] Olivier Debeir, Isabelle Van Den Steen, Patrice Latinne, Philippe Van Ham, and Eleonore Wolff. Textural and contextual land-cover classification using single and multiple classifier systems. *Photogrammetric Engineering and Remote Sensing*, 68: 597–605, 2002.
- [196] D. Lu and Q. Weng. A survey of image classification methods and techniques for improving classification performance. *Int. Journal of Remote Sensing*, 2007.
- [197] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L. Littman. Activity recognition from accelerometer data. In *Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence*, pages 1541–1546, 2005. ISBN 1-57735-236-x.
- [198] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. ISSN 1573-0565.
- [199] Marina Skurichina and Robert P.W. Duin. Bagging for linear classifiers. *Pattern Recognition*, 31(7):909 – 930, 1998. ISSN 0031-3203.
- [200] Rafael Cruz, Robert Sabourin, and George Cavalcanti. Dynamic classifier selection: Recent advances and perspectives. *Information Fusion*, 41, 05 2018.
- [201] Rafael Cruz, Robert Sabourin, George Cavalcanti, and Tsang Ing Ren. Meta-des: A

- dynamic ensemble selection framework using meta-learning. *Pattern Recognition*, 48, 05 2015.
- [202] Paulo Rodrigo Cavalin, Robert Sabourin, and Ching Y. Suen. Dynamic selection approaches for multiple classifier systems. *Neural Computing and Applications*, 22: 673–688, 2011.
- [203] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification, 06 2010.
- [204] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomput.*, 174(PA):50–59, January 2016. ISSN 0925-2312.
- [205] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. Multimodal neural language models. In *Proceedings of the 31st International Conf. on ML*, volume 32, 22–24 Jun 2014.
- [206] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML’11*, 2011.
- [207] Tom Zahavy, Abhinandan Krishnan, Alessandro Magnani, and Shie Mannor. Is a picture worth a thousand words? A deep multi-modal architecture for product classification in e-commerce. In *AAAI*. AAAI Press, 2018.
- [208] Corey Lynch, Kamelia Aryafar, and Josh Attenberg. Images don’t lie: Transferring deep visual semantic features to large-scale multimodal learning to rank. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pages 541–548, 2016. ISBN 978-1-4503-4232-2.
- [209] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo col-

- lections. In *Computer Vision – ECCV 2014*, pages 529–545, 2014. ISBN 978-3-319-10593-2.
- [210] A. Kannan, P. P. Talukdar, N. Rasiwasia, and Q. Ke. Improving product classification using images. In *2011 IEEE 11th Int. Conf. on Data Mining*, Dec 2011.
- [211] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129. 2013.
- [212] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *IEEE CVPR*, 2016.
- [213] L. Wang, Zhe Wang, Wenbin Du, and Yu Qiao. Object-scene convolutional neural networks for event recognition in images. *CoRR*, abs/1505.00296, 2015.
- [214] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT’ 98, pages 92–100, New York, NY, USA, 1998. ACM. ISBN 1-58113-057-0.
- [215] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [216] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

- [217] Mark Laan, Eric C Polley, and Alan Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6:Article 25, 02 2007.
- [218] Leo Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, 1996.