Testing bias in analysis of convolutional neural networks

by

Sanchari Dhar

B.Tech, RCC Institute Of Information Technology, 2016

A THESIS

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Approved by:

Major Professor
Dr Lior Shamir

Copyright

Abstract

Deep convolution neural networks (DCNNs) have become extremely common in computer vision, and due to the availability of easy-to-use libraries, their impact has gone far beyond the domain of computer vision. Since a DCNN acts like a black box, it is very often difficult for the user to understand which features of the image contribute to the learning of the network. The purpose of this work is to explore the reliability of DCNNs as general solutions to machine vision problems and identify possible weaknesses in which DCNNs can lead to biased or misleading results. A first experiment shows that for a basic classification of spiral and elliptical galaxies, the position of the galaxies plays role in the classification. That small but consistent and statistically significant bias can lead to misleading results when applied to large datasets. The second experiment has been done with a variety of prominent datasets in the computer vision domain. Only a portion of the background without any significant content descriptor has been used, but still, the LeNet5 architecture is able to predict the image better than the mere chance accuracy. That shows that the classification accuracy, even when using commonly used datasets, can be biased.

Table of Contents

List of Figures

List of Tables

Acknowledgments

Foremost, I would like to convey my sincere gratitude to my advisor Dr. Lior Shamir, for his unwavering confidence and trust in my MS research and thesis. I am highly grateful for Dr. Shamir's patience, his constant motivation, his expertise, and most importantly his guidance which was the key for the completion of this MS thesis project.

I would also like to thank my MS committee members : Dr. Pascal Hitzler, and Dr. Torben Amtoft for their reviews, suggestions, and inspiration.

I would like to thank my parents, Chandan Dhar and Sampa Dhar, and family Ranojoy for their guidance and support towards me, their unconditional love helped me achieve everything I have today.

I would also like to thank my friends Saptarshi, Meghna and Abhilekha for their unwavering believe in me during this period. Thank you everyone for your stimulating discussions and support.

Finally, I would like to thank god for showering me with his grace and blessing.

Thank you everyone.

Introduction

Deep Learning is a machine learning technique that is inspired by the structure of the human brain. Convolution neural networks(CNN) is arguably the most popular deep learning architecture. It is a key technology in various domains such as driverless cars, recognize a stop sign, or differentiating between a pedestrian from a lamppost.[1]. It is also successfully applied to recommender systems, natural language processing, automated image and text interpretation in medicine, insurance, advertisement, public video surveillance, job applications, or credit scoring, and many more.[1] CNNs have an edge over other primitive machine learning models because it does not require any human intervention to detect important features of the image in the dataset while learning. [2]

A CNN architecture is composed of multiple convolution layers stacked together. The main part of the architecture is an input layer, an output layer, and many hidden layers in between. In the hidden layers, a convolution filter is applied to the input data to extract a feature map which is fed into the output layer to wrap up the working of the convolution neural network. This complex process of distilling the ordinal features from the image helps the model to perform wonderful feats, but this complexity is also a curse. The non-intuitive rules of the hidden layers is often a mystery to the creator and make the network a "Black Box" [3]

According to past research on CNN architecture, it has been shown that CNN sometimes produces sub-optimal results in comparisons to existing ML techniques. Research carried out by[2][3] have shown that output of CNN is not superior than result of sentence weight neural network for detecting spam reviews. Furthermore, it also has been shown that the accuracy of CNN is lower than the accuracy of a bidirectional encoder-decoder[4] when considered for

[1]https://www.mathworks.com/discovery/deep-learning.html

[2]https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

[3]https://bdtechtalks.com/2021/01/11/concept-whitening-interpretable-neural-networks/

classification type of problems. In the paper by Zhang et al[5] have linked the repetitive features to be semantically linked to the target attributes. The co appearing features in the dataset imparts a bias to the target values.

Scientists have been researching the weakness of Deep learning models even when the models are performing with high accuracy. For instance, a good model can totally extract wrong features to make a good decision because the CNN learns from the training data to characterize the task.[1] For example if the model needs to classify between a dog and a wolf it turns out that the classifier can able to detect the snow in the background of the wolf more precisely than the actual important features that should be used by the model to classify. Though the model gives a high accuracy rate the reason is the bias in the background of the training set.[1;6] Similarly a classifier used to differentiate between the enemy tank and friendly tank delivers a high accuracy of classification but turns out to be a good classifier of sunny or overcast days.[1;7] However this example shows how harmful it could be to depend on the results of a black box with high performing accuracy.

In this paper we study astronomical datasets which uses supervised learning process for a basic classification between spiral and elliptical morphology. Though the classification accuracy is above 90% but there is a subtle but consistent bias in the dataset. The different part of the sky in the training set lead to a consistent bias in the classifier based on the sky location of the galaxies it attempts to classify. If this bias is ignored it does not reflect the actual distribution of the morphology in the sky. We also study other several benchmark dataset in the field of face recognition or biomedical image recognition and object recognition to find that the classification accuracy of this datasets are driven by dataset bias due to the data acquisition process and hence lead to false experimental result.

We recommend simple possible solution that can be completed before making any closure about a classification problem using convolution neural networks.

Chapter 1

The risk in using deep neural networks for annotating large catalogs of astronomical images

Abstract

Deep convolutional neural networks (DCNNs) have become the most common solution for automatic image annotation due to their non-parametric nature, good performance, and their accessibility through libraries such as TensorFlow. Among other fields, DCNNs is also a common approach to the annotation of large astronomical image databases acquired by digital sky surveys. One of the main downsides of DCNNs is the complex non-intuitive rules that make DCNNs act as a "black box", providing annotations in a manner that is unclear to the user. Therefore, the user is often not able to know what information is used by the DCNNs for the classification. Here we demonstrate that the training of a DCNNs sensitive to the context of the training data such as the location of the objects in the sky. We show that for the basic classification of elliptical and spiral galaxies, the sky location of the galaxies used for training affects the behavior of the algorithm, and leads to a small but consistent and statistically significant bias. That bias exhibits itself in the form of

cosmological-scale anisotropy in the distribution of basic galaxy morphology. Therefore, while DCNNs are powerful tools for annotating images of extended sources, the construction of training sets for galaxy morphology should take into consideration more aspects than the visual appearance of the object. In any case, catalogs created with deep neural networks that exhibit signs of cosmological anisotropy should be interpreted with the possibility of consistent bias.

1.1 Introduction

In the past two decades, autonomous digital sky surveys powered by robotic telescopes have been becoming increasingly important in astronomy, and have been revolutionizing astronomy research. The ability to collect far more data than any manually controlled telescope and make them accessible to the community through the concept of virtual observatory increased the efficiency of telescope systems, leading to unprecedented discovery power[8]. It also allows making the astronomy research community broader, as any person with a computer and connection to the Internet can gain immediate access to powerful astronomical research instruments.

Perhaps the first major comprehensive autonomous digital sky survey is the Sloan Digital Sky Survey[9], with considerable success and revolutionary impact on astronomy. The overwhelming success of SDSS was followed by other powerful sky surveys such as the Panoramic Survey Telescope and Rapid Response System[10] and the Dark Energy Survey[11]. Future ventures such as the Vera Rubin Observatory and the space-based Euclid mission will provide even more powerful imaging capabilities, leading to far greater databases and consequently discovery power.

As digital sky surveys image hundreds of millions and even billions of astronomical objects, it is clear that manual analysis of the data is impractical. One of the most challenging tasks in the analysis of the image data acquired by digital sky surveys is the morphological analysis of extended objects. Unlike point sources, extended objects can have a complex morphology, and their analysis requires sophisticated computational methods. Tasks related

to automatic annotations of galaxies can include broad classification of galaxies to elliptical or spiral[12–15], or annotation of a more comprehensive set of morphological descriptors of galaxies[16;17]. Other tasks can include automatic detection of rare galaxies[18–21], unsupervised analysis of galaxy morphology[22], or separating galaxies from stars[23].

In the past decade, deep convolutional neural networks (DCNNs) have been becoming increasingly more common in machine vision. Their good performance combined with their non-parametric approach and the availability of open-source libraries makes DCNNs an effective solution that allows achieving good performance, yet with reasonable development efforts. As they become popular in almost all fields that involve machine vision, DCNNs have also been becoming very common in astronomy. Among other tasks, they are also used for automatic annotation of galaxy images[12;17;24–27]. Due to their efficiency and speed, DCNNs are currently the immediate solution for the annotation of very large datasets of galaxy images.

However, while DCNNs have the important advantages mentioned above, they also have several weaknesses. One of the main downsides of DCNNs is the "black box" nature of its classification process. DCNNs are trained by data samples, and the weights are determined during training to optimize the performance. However, the rules by which the classifications are made are complex and non-intuitive, making it difficult to conceptualize the way the classifications are being made by the neural network. Since it is difficult to define what the DCNNs "learn" from the data, such systems should be used with caution[28;29].

Here we demonstrate that a DCNN system used through a typical supervised machine learning process to identify galaxy morphology can provide good accuracy in the classification of the galaxies, but at the same time can have a subtle but consistent bias. When applied to large datasets, that bias can lead to consistently biased catalogs. If the bias is ignored in the consequent analysis of the catalog, it can lead to observations that do not reflect the distribution of the data in the real sky.

1.2 Method

We train a neural network to classify between elliptical and spiral galaxies. The distribution of elliptical and spiral galaxies in one part of the sky is expected to be statistically the same as the distribution of elliptical and spiral galaxies in other parts of the sky. That is, when the dataset of galaxies is large, the shape of elliptical galaxies observed in a certain RA and declination range is expected to be the same as the shape of elliptical galaxies in any other RA and declination range. In other words, an expert observing an image of a random elliptical galaxy, and given no other information about the object, will not be able to make a knowledgeable guess of the RA and declination of that galaxy.

According to the null hypothesis, the neural network is trained by the morphology of the galaxies, and therefore the classification output of the neural network depends only on the morphology of the galaxy. In that case, if the deep neural network is trained with a high number of galaxies, it will perform the same way (within statistical error) regardless of the location of the test galaxy in the sky. Otherwise, the neural network is sensitive also to the location of the galaxy in the sky and therefore can lead to a certain bias. Even if such bias is small when annotating a very large number of galaxies that bias can be statistically significant. For instance, if in a certain part of the sky the neural network tends to classify more galaxies as the spiral, a catalog generated by that network will show cosmological anisotropy such that a certain direction of observation has more spiral galaxies compared to other directions of observation.

In this study, we train a deep neural network to classify between elliptical and spiral galaxies by using training galaxies imaged in the same part of the sky. We then compared the results using the same test set, but such that the training spiral galaxies are taken from one part of the sky, and the test spiral galaxies are taken from another part of the sky. If the neural network classifies galaxies just by their morphology, both neural networks should provide the same confusion matrix, within statistical error. However, if the confusion matrices are different, it means that the neural network also learns differences in the sky background, and can be affected based on the location of the galaxy in the sky. When

applied to a large dataset, such behavior of the neural network can lead to differences in the distribution of the annotations based on different parts of the sky. That can consequently lead to slightly but consistently biased data products.

1.2.1 Deep convolutional neural network

The deep convolutional neural network used in this experiment is an expansion of the common LeNet-5 architecture[30], implemented using the Keras library[31;32], and adjusted to the input size of images with the dimensionality of 120×120 pixels. The deep neural network model is based on the sequential model of Keras with five convolution layers and four max-pooling layers. These layers are followed by flattening and fully connected layers.

The activation function used in most layers of the convolutional neural network is Rectified Linear Unit (ReLU), except for the output layer, where we use the sigmoid activation function. Figure 1.1 shows the diagram of the structure of the convolutional neural network. During compilation, the model uses the Adam (Adaptive Moment Estimation) optimizer[33], with an adaptive learning rate, and the binary cross-entropy is used as the loss function because of binary classification.

1.3 Data

Datasets from two major digital sky surveys were used - SDSS and Pan-STARRS. To have the training and test sets of spiral galaxies, we used catalogs of galaxies annotated by their broad morphology to spiral and elliptical galaxies. The image data in both datasets are the 120×120 JPG images downloaded by using the *cutout* service.

For SDSS, the annotation of spiral and elliptical galaxies were taken from a catalog of galaxies annotated by their broad morphology[34;35]. Each galaxy in the catalog is provided with its annotation, and the certainty of the annotation in the range (0.5,1), where 1 is the maximum certainty for the galaxy to belong in the morphological type it is annotated. To ensure the accuracy of the annotations, only galaxies with a certain threshold of 0.9 or

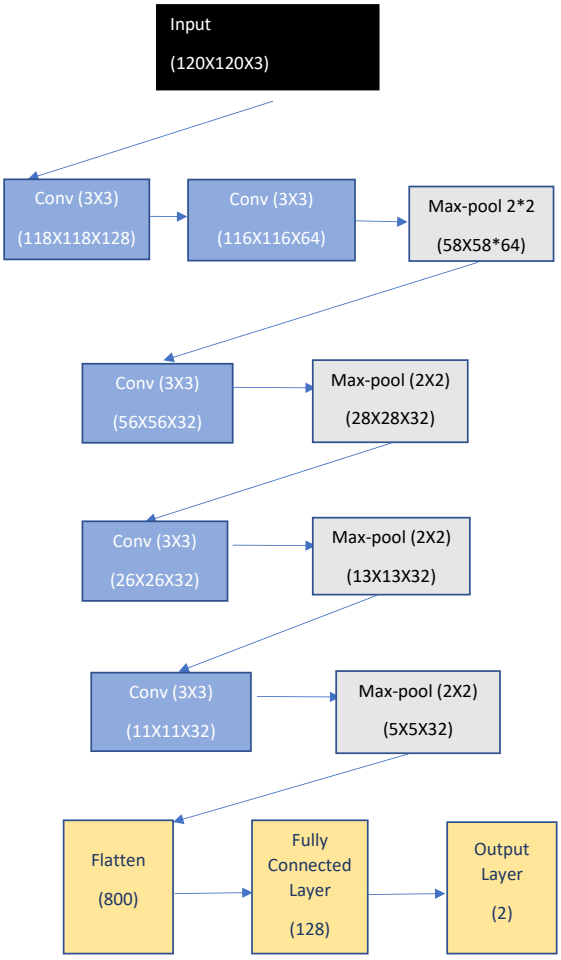

Figure 1.1: *Diagram of the structure of the convolutional neural network.*

higher were used. These galaxies have certainty of their annotation of ∼98% compared to the "superclean" Galaxy Zoo annotations[34]. The total number of galaxies in the catalog is $\sim 2.9 \cdot 10^6$.

A similar catalog was also used For the Pan-STARRS data. The catalog of Pan-STARRS galaxies contained $\sim 1.7 \cdot 10^6$ automatically annotated galaxies imaged by Pan-STARRS[36]. Like with the SDSS galaxies, only galaxies with annotation certainty of 90% or higher were used, to ensure that the dataset for training and testing the neural network is clean.

For Pan-STARRS data, datasets of galaxies were taken from two opposite hemispheres in the sky. The RA ranges of the sky regions used in the experiments are $(0^o - 20^o)$, and $(180^o - 200^o)$. The declination in both cases is $(0^o - 20^o)$. These regions were chosen for

being far from each other in the sky, but also because they contain a sufficient number of galaxies. Additionally, a dataset of Pan-STARRS galaxies in the RA range $(180^o - 200^o)$ and declination range of $(20^o - 40^o)$ is also used. The sky regions used in the SDSS data are RA ranges $(230^o - 260^o)$ and $(15^o - 35^o)$ for the RA. The declination range is $(-10^o - 40^o)$.

Table 1.1 summarizes the number of objects taken from each sky region in each sky survey, and the number of objects used for testing and training. Naturally, the training and test sets are completely orthogonal, and no galaxies can be assigned for training in one experiment, and for testing in another experiment.

Dataset	Sky Survey	RA range	Dec range	Train Elliptical	Train Spiral	Test Elliptical	Test Spiral
1	Pan-STARRS	0^o-20^o	0^o-20^o	3000	3000	8000	8000
2	Pan-STARRS	180^o-200^o	0^o-20^o	3000	3000	8000	8000
3	Pan-STARRS	180^o-200^o	20^o-40^o	3000	3000	8000	8000
A	SDSS	230^o-260^o	-10^o- 40^o	2000	2000	3000	3000
B	SDSS	15^o-35^o	-10^o- 40^o	2000	2000	3000	3000

Table 1.1: *The datasets used in the experiment, and the number of training and test galaxies in each dataset. RA and declination ranges are in degrees.*

1.4 Results

The DCNN method described in Section 1.2.1 was tested with different combinations of the datasets described in Section 1.3. For the baseline experiment, we trained and tested the model with the galaxies in the same right ascension and declination ranges. Table 1.2 shows the confusion matrix when classifying the Pan-STARSS galaxies using the spiral and elliptical galaxies from Dataset 1 for both training and test purpose. The number of training and test galaxies are specified in Table 1.1.

	Elliptical	Spiral
Elliptical	7850	150
Spiral	756	7244

Table 1.2: *Confusion matrix of the classifications when using Dataset 1 for both training and testing.*

As the table shows, more spiral galaxies were incorrectly classified as elliptical galaxies compared to elliptical galaxies classified incorrectly as a spiral. Therefore, if applied to a

very large dataset of galaxies, it will show a slightly higher fraction of elliptical galaxies than spiral galaxies. However, a slight bias is expected from any classifier, and such bias is not necessarily expected to lead to an observation of large-scale anisotropy.

The reason a bias in the classifier is not expected to lead to an observation of anisotropy is that such bias is expected in all parts of the sky. That is, regardless of the location of the test galaxies in the sky, a higher number of elliptical galaxies is expected, and therefore the ratio between elliptical and spiral galaxies is not expected to change throughout the sky.

Since the number of elliptical and spiral galaxies is certainly not expected to be equal, and since the separation between spiral and elliptical galaxies is not strictly due to the many in-between cases, a slight but consistent bias of the algorithm might not necessarily lead to false large-scale anisotropy. Certainly, the performance of such algorithms also depends on the imaging, as higher resolution images allow to better identify spiral features of galaxies, increasing the number of spiral galaxies compared to elliptical galaxies[36]. If spiral features are identified, that indicates that the galaxy is indeed spiral. However, if spiral features are not identified it could also be because the image resolution does not allow the identification of the spirality of the galaxy[36].

The same analysis was also done using Dataset 2. Table 1.3 shows the confusion matrix when classifying the Pan-STARSS galaxies using Dataset 2 for both training and testing.

	Elliptical	Spiral
Elliptical	7699	301
Spiral	450	7550

Table 1.3: *Confusion matrix of the classification when using Dataset 2 for both training and testing.*

The error rate for the spiral and elliptical galaxies is somewhat different from the error rate shown in Table 1.2, and the difference provides a certain indication of a link between the performance of the classifier and the part of the sky from which the galaxies are taken. However, in these two experiments, the training and test data are different, and therefore no conclusive evidence of a link between a consistent bias of the classifier and the part of the sky from which the training data are taken can be inferred.

To further investigate a possible link between the part of the sky from which the training data are taken and a consistent bias of the classifier, the training set was designed such that the spiral galaxies were taken from Dataset 1, and the elliptical galaxies from Dataset 2. The test galaxies are all from Dataset 1. Table 1.4 shows the confusion matrix of the predictions made by the classifier.

	Elliptical	Spiral
Elliptical	7749	251
Spiral	408	7592

Table 1.4: *Confusion matrix of the classifications when using spiral galaxies from Dataset 1, elliptical galaxies Dataset 2 for training, and all test galaxies from Dataset 1 for testing.*

As the table shows, although the test galaxies are the same, the distribution of the galaxies is different from the results of the experiment shown in Table 1.2. According to Table 1.4, more spiral galaxies are classified correctly, while more elliptical galaxies are classified incorrectly.

According to Table 1.4, ∼49.02% are predicted as a spiral, and therefore the probability of a galaxy to be predicted as the spiral in that dataset is 0.4902. According to the results of Table 1.2, merely 7,394 galaxies were classified as a spiral. According to the binomial distribution, if the success probability is 0.4902, the probability of having 7,394 or less successful events is $P < 10^{-5}$. That shows that although the test set is identical, the predictions are significantly different when using training data from different parts of the sky.

The high probability shows that if the classifier was applied to galaxy images that are not labeled with ground truth, it would have shown a statistically significant difference between the frequency of spiral galaxies in the sky region of $(0^o < \alpha < 20^o, 0^o < \delta < 20^o)$ and the frequency of spiral galaxies in the sky region $(180^o < \alpha < 200^o, 0^o < \delta < 20^o)$. These differences could be interpreted as evidence for cosmological-scale anisotropy.

To further examine a possible link between the selection of the training data and the behavior of the classifier, a similar experiment was performed such that the training data was spiral galaxies from Dataset 1 and elliptical galaxies from Dataset 2. Then, the performance

of the classifier was tested with test data from Dataset 2. Table 1.5 shows the confusion matrix of the experiment.

	Elliptical	Spiral
Elliptical	7791	209
Spiral	462	7538

Table 1.5: *Confusion matrix of the classifications when using spiral galaxies from Dataset 1, elliptical galaxies from Dataset 2 for training, and Dataset 2 used for testing.*

As the table shows, when compared to the classifications of the same set of galaxies with a neural network that was trained with galaxies from the same sky region as the test data, the number of misclassified spiral galaxies increases from 450 to 462, and the number of misclassified elliptical galaxies decreases from 301 to 209. If the probability of misclassified elliptical galaxy is $\frac{301}{8000} = 0.037625$, the probability of having 209 or fewer misclassified elliptical galaxies from the 8,000 galaxies that were classified is ($P < 10^{-5}$). Because the galaxies have ground-truth we can be certain that the reason for the difference is not of astronomical origin, but the higher similarity of the test elliptical galaxies from Dataset 2 and the training elliptical galaxies from Dataset 2, which leads to a consistent bias in the classification.

In another experiment, the training set contained spiral galaxies from Dataset 2 and elliptical galaxies from Dataset 1. Tables 1.12 and 1.7 show the confusion matrix when applying the classifier to the test samples of Dataset 1 and Dataset 2, respectively.

	Elliptical	Spiral
Elliptical	7592	408
Spiral	551	7449

Table 1.6: *Confusion matrix of the classification when training the neural network with spiral galaxies from Dataset 2, elliptical galaxies from Dataset 1, and testing the classifier with spiral and elliptical galaxies from Dataset 1.*

The results show that although the training set is the same in both cases, each dataset provided different results. For the test samples of Dataset 1, a higher number of spiral galaxies was misclassified as elliptical galaxies, while when classifying the test samples of Dataset 2 more elliptical galaxies were classified as spiral galaxies. That is, the classifier

	Elliptical	Spiral
Elliptical	7514	486
Spiral	342	7658

Table 1.7: *Confusion matrix of the classification when the spiral galaxies of the training set are from Dataset 2, the elliptical galaxies of the training set are from Dataset 1, and the test samples are the test galaxies from Dataset 2.*

showed 7,857 spiral galaxies in Dataset 1, and 8,144 spiral galaxies in Dataset 2. Because the galaxies are annotated with ground truth, we can conclude that the reason for the difference is not an actual higher number of spiral galaxies in the real sky at ($180^o < \alpha < 200^o, 0^o < \delta < 20^o$), but a higher similarity of the galaxies in the test set to the galaxies in the training set that were taken from the same part of the sky. However, if the galaxies did not have ground truth, the difference could have been interpreted as an indication of cosmological-scale anisotropy.

1.4.1 Difference between close sky regions

The experiments above tested for the impact when the different classes in the training set are imaged in opposite hemispheres. To test whether the same bias also occurs when the classes of the training data are acquired in closer regions, another experiment was done such that the different sky regions are neighboring.

Table 1.8 shows the confusion matrix when the neural network was trained with elliptical galaxies from Dataset 2, and spiral galaxies from Dataset 3. Dataset 2 was used for testing. These results are compared to the confusion matrix of Table 1.9, showing the classifications when the same test set was used, but the neural network was trained with elliptical galaxies from Dataset 3, and spiral galaxies from Dataset 2. The two confusion matrices show some differences in the classifications, but less substantial than when the training set was taken from different parts of the sky in opposite hemispheres.

	Elliptical	Spiral
Elliptical	7755	245
Spiral	600	7400

Table 1.8: *The confusion matrix when the classifier is trained with elliptical galaxies from Dataset 2, and spiral galaxies from Dataset 3. The test set is the test galaxies from Dataet 2.*

	Elliptical	Spiral
Elliptical	7702	298
Spiral	516	7484

Table 1.9: *The confusion matrix when the neural network is trained with elliptical galaxies from Dataset 3, and spiral galaxies from Dataset 2. The test set is the test galaxies from Dataet 2.*

1.4.2 Experiments with SDSS data

In addition to the Pan-STARRS data, we also tested data from the Sloan Digital Sky Survey (SDSS). Table 1.10 shows the confusion matrix of the classifications of the test galaxies of Dataset A when the classifier was trained with training spiral galaxies from Dataset A, and training elliptical galaxies from Dataset B. As before, no galaxies were included in both the training and test sets.

	Elliptical	Spiral
Elliptical	2704	296
Spiral	31	2969

Table 1.10: *Confusion matrix of the classification of SDSS dataset when training the convolutional neural network with spiral galaxies from Dataset A and elliptical galaxies from Dataset B. The test samples are from Dataset A.*

Table 1.11 shows the confusion matrix when using the same training set as was used for the experiment shown in Table 1.10. As the table shows, the number of misclassified spiral galaxies increases to 85, while the number of misclassified elliptical galaxies drops to 109. If the probability of an elliptical galaxy to be misclassified as the spiral galaxy is $\sim$0.0987, the binomial distribution probability to have 109 or fewer misclassified elliptical galaxies is $P < 10^{-5}$.

Like with the Pan-STARRS data, the analysis shows that the classifier shows the different ratios of elliptical and spiral galaxies even when the test set is identical. Because the test

samples are the same, and the algorithm is the same, the only possible explanation for the difference is the use of a different training set. For example, the increase in the number of galaxies classified as elliptical galaxies can be linked to the fact that the neural network was trained with elliptical galaxies imaged in the same part of the sky of the test samples, while the training spiral galaxies were imaged in a different part of the sky. The bias is statistically significant. Therefore, using that neural network to classify galaxies in the entire sky would lead to a statistically significant difference between the frequency of spiral galaxies in different parts of the sky, which might provide evidence of cosmological-scale anisotropy.

	Elliptical	Spiral
Elliptical	2891	109
Spiral	85	2915

Table 1.11: *Confusion matrix of the classification of SDSS galaxies when training the neural network with spiral galaxies from Dataset A and elliptical galaxies from Dataset B, and test set is the test samples of Dataset B.*

A similar experiment was done such that the training set was made of spiral galaxies from Dataset B and elliptical galaxies from Dataset A. Tables 1.12 and 1.13 show the confusion matrices when testing the neural network with test data from Dataset A and Dataset B, respectively. As the confusion matrices show, the bias identified in the results is consistent with the results of the previous experiment.

Although the classifier is the same classifier trained with the same data, the misclassifications of Dataset A are completely different than the misclassifications of Dataset B. When testing the classifier with the test data of Dataset A, much more galaxies are classified as elliptical compared to the confusion matrix produced when the classifier was tested with Dataset B. Given that the probability of a test spiral galaxy in Dataset A to be misclassified as an elliptical galaxy is 0.165, the probability to have 154 misclassified galaxies or less is $P < 10^{-5}$.

	Elliptical	Spiral
Elliptical	2933	67
Spiral	495	2505

Table 1.12: *Confusion matrix of the classifications of SDSS galaxies when training the neural network with spiral galaxies from Dataset B and elliptical galaxies from Dataset A. The test set is spiral and elliptical galaxies from the test samples of Dataset A.*

	Elliptical	Spiral
Elliptical	2837	163
Spiral	154	2846

Table 1.13: *Confusion matrix of the classifications of SDSS galaxies when training the neural network with spiral galaxies from Dataset B and elliptical galaxies from Dataset A. The test samples are the test spiral and elliptical galaxies from Dataset B.*

1.5 Conclusion

As autonomous digital sky surveys generate vast pipelines of image data, including billions of extended objects with complex morphology, a solid approach to analyze the morphology of these objects is by applying deep convolutional neural networks. For the purpose of supervised machine learning, these networks are trained automatically with labeled "ground truth" samples, and can then annotate any given new data based on the rules deduced in the training stage.

The application of convolutional neural networks to large image databases collected by digital sky surveys can produce large catalogs of annotated objects. It is expected that these data products will be used by other researchers to answer questions that were difficult to address observationally in the pre-information era, such as the large-scale structure of the universe.

However, while deep neural networks can provide fast annotation with a high level of accuracy, they are based on complex and non-intuitive data-driven rules that are difficult to interpret and fully understand. Therefore, these rules can reflect not just the morphology of the galaxy, but in fact any piece of information by which the neural network can differentiate between the different classes of images it is trained with.

Here we show that while deep convolutional neural networks provide good annotation

accuracy, the training process can potentially introduce subtle but consistent biases. Namely, we show that unbalanced distribution of the sky location of the galaxies in the training set can lead to a consistent bias of the classifier, which can lead to a bias in the classifier based on the sky location of the galaxies it attempts to classify. When applied to very large databases typical to astronomical sky surveys, even a small bias can become statistically significant, and might even mislead potential users of data products generated by deep neural networks into false conclusions.

The examples shown in this paper are focused on specific datasets and annotation tasks, and it is very reasonable to assume that many systems based on deep neural networks are not biased. However, these examples demonstrate that such systematic bias can exist, and should be taken into considerations when designing neural networks for annotation of astronomical images, and when using data products generated by these neural networks. Cosmic variance, different atmospheric conditions, and even different states of the hardware when training data are acquired can affect the training of an artificial neural network, and allow the network to learn non-astronomical information that can differentiate between the classes in the training set.

Deep convolutional neural networks have become increasingly more common in astronomy and can be used for a broad range of tasks. However, as such neural networks heavily rely on training data, it is difficult to acquire training data that evenly covers all parts of the sky under all weather conditions and status of the hardware. Therefore, when analyzing data using deep neural networks, a certain bias is expected. The use of data products produced by these networks should therefore be used with consideration of the advantages as well as the disadvantages of deep neural networks and should be matched to tasks for which such data annotation process is scientifically sound.

Acknowledgment

The research was supported by NSF grant AST-1903823. The Pan-STARRS1 Surveys (PS1) and the PS1 public science archive have been made possible through contributions by the

Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max-Planck Society, and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg and the Max Planck Institute for Extraterrestrial Physics, Garching, The Johns Hopkins University, Durham University, the University of Edinburgh, the Queen's University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, the National Aeronautics and Space Administration under Grant No. NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation Grant No. AST-1238877, the University of Maryland, Eotvos Lorand University (ELTE), the Los Alamos National Laboratory, and the Gordon and Betty Moore Foundation.

Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions.

SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, Center for Astrophysics — Harvard & Smithsonian, the Chilean Participation Group, the French Participation Group, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt

University, and Yale University.

Chapter 2

Evaulation of the reliability of deep convolutional neural network fordata science

2.1 Abstract

Convolutional neural networks (CNNs) have been becoming a key paradigm used in data science with image data. CNN provides superior performance yet without requiring the user to develop specific algorithms. With the availability of easy-to-use libraries, CNNs have become the default solution for image classification problems in data science. However, one of the disadvantages of CNNs is the complex non-intuitive rules that can often be considered"black box". Therefore, CNNs are vulnerable to learning from irrelevant information in the datasets they are trained with. Here we test the reliability of CNNs in image classification problems in several different tasks. The experiments show that CNNs can provide promising classification accuracy even when they are trained with a dataset that does not contain any relevant information, or can be systematically biased by irrelevant information in the image data. The presence of such consistent irrelevant data is difficult to identify and can therefore lead to false or biased experimental results. Possible solutions to this downside

of CNNscan be control experiments, as well as other protective practices to validate the results and avoid biased conclusions based on CNN-generated annotations.

2.2 Introduction

Automatic analysis of image data has been becoming an important part of data science. Motivated by the increasing availability of digital imaging and large storage devices, the ability to analyze large databases of images is now pivotal in discovery from data in a broad range of domains.

Convolution Neural Networks (CNNs) have become a primary tool in the analysis of image data, making automatic image analysis with machine learning more powerful and more accessible. CNN's allow learning directly from the pixels, and therefore can be used for a very broad range of image analysis problems yet without the need to develop specific algorithms. The ability to analyze different types of images without designing specific algorithms, combined with superior performance, made CNN the default solution for automatic analysis of image data. With the availability of open-source libraries, CNNs have been made accessible to users who are not necessarily machine learning experts, expanding the use of CNN's far beyond the machine learning community.

By learning directly from the pixels, CNNs identify complex non-intuitive features that can be used to identify the images, and in the case of supervised machine learning can associate each image with its class. However, learning from the pixels without the need to design specific image content descriptors rewards the CNN for using any possible information in the visual content that can discriminate between the different image classes. That can include information that is relevant to the image classification problem at hand, but also information that might not necessarily be driven by the intended content of the images. For instance, pixels might have different values based on subtle differences in the lighting conditions at the time of imaging, slight changes in the position of the camera or subject, and even differences in the temperature of the CCD when the images are taken.

Benchmark datasets can be biased for different reasons. For instance, in the context of

benchmarks for object recognition, the perception of the people annotating the samples or selecting the samples for the dataset can lead to bias, especially when the dataset is collected from the web[37–40]. That bias can be shown experimentally by the fact that training an algorithm with one benchmark dataset and testing it with another dataset leads to weaker results compared to training and testing using the same benchmark[37]. That difference in performance is counter-intuitive, especially give that larger training sets are expected to provide stronger or equal performance to smaller training sets, and therefore the weaker performance can be considered evidence to dataset bias[37].

One of the solutions to the problem of dataset bias is to increase the variability in the datasets. That can be done by reducing the dataset bias by using data augmentation[41;42], combining different datasets[38], or synthetically change the variability of the dataset[38].

Benchmark datasets used in the domain of machine learning aim at representing the real world as reliably as possible[37], allowing to development and compare the performance of different algorithms that solve general common problems such as automatic object recognition or face recognition. However, in data science datasets are often collected for one single experiment, and used by a single research team. That might make it more difficult to identify and characterize all possible biases in the dataset. While substantial work has been done to analyze bias in datasets that were collected from the web, here we focus on biases in datasets collected in a controlled environment and well-defined data acquisition processes.

Assuming no bias in the dataset, the performance of a convolutional neural network achieved when analyzing that dataset can be trusted as an indication of the presence of a signal in the dataset. The application of the trained CNN to large datasets can be used to make a discovery in the data. However, if the dataset is biased in a certain way, that bias could lead to signals driven by the bias rather than by the actual data. That misleading situation can be carried on to the application of the CNN to the analysis of the large dataset, and consequently to false discoveries that are consistent and statistically significant but are driven by dataset bias.

A controlled data acquisition process does not necessarily guarantee unbiased datasets. For instance, a medical dataset for automatic image-based diagnostics can be acquired at

more than one clinic. If the positive cases are not distributed equally across the different clinics, a CNN can learn the features that characterize a certain clinic, and in fact, develop an algorithm for "clinic prediction" rather than identification of the actual medical condition being analyzed. Because different clinics can use different hardware, different settings, and different technicians, it is extremely difficult to guarantee that the image acquisition process in all clinics is identical.

An example of such bias was demonstrated using microscopy data, where an algorithm could predict the treatment applied to cells in microscopy images[43]. But the ability to classify the cells was also driven by the imaging session rather than the morphology of the cells[44]. That was shown by the consistency of the results regardless of the presence of cells in the images, demonstrating that the signal was driven by the background noise rather than the cells[44]. In many cases the visual features leading to the bias are too subtle to notice by eye, making it difficult to notice the presence of the bias, and leading the experimentalist to believe that the predictions made by the CNN reflect the ability of the CNN to identify differences between the visual content. Consequently, the experimentalist might reach certain conclusions regarding the differences between the image classes.

Here we study several different datasets acquired in a controlled process. That includes benchmarks used for common tasks such as object recognition and face recognition, but also datasets used for discovery-driven research in the medical and astronomical domains. We show that in many cases the application of the CNNs to the datasets leads to results-driven by dataset bias related to the data acquisition process, and can therefore lead to false discoveries. To avoid biased results, we propose simple control experiments that should be performed before making conclusions from the result of the application of convolutional neural networks to image data.

2.2.1 Medical images

As an example of a dataset of biomedical images, the dataset COVID-CT was used. Details about the dataset are shown in Table 2.1. The dataset was used for CoroNet[45], a deep

convolution neural network that can identify COVID-19 infection from chest X-ray images. *COVID-CT* has two classes such that each class has 349 images for both training and testing purposes.

Figure 2.3 shows examples of original chest x-rays images from the dataset, and a 20×20 pixels sub-images from the top left corner of the original image. The human eye is not able to identify differences between the different classes based on the cropped sub-images alone, as these are blank background area that does not contain images of any part of the body. As Table 2.1 shows, despite the absence of information related to COVID in the seemingly blank background subsets, the DCNN was able to achieve classification accuracy much higher than mere chance.

Classification accuracy of ∼62.5% was observed with the original dataset for *COVID-CT* when classifying the images into COVID or non-COVID using LeNet-5 architecture. This shows that the dataset of the sub-images provided better prediction than the original dataset, which can be due to the signal from the differences in the imaging process is stronger than the signal from the medical condition reflected by the images. A dataset of subset images is more consistent, allowing the DCNN to learn the subtle but consistent differences between the images originated from the imaging process. Another possible reason could be due to the reduction in the size of the image from 311×224 pixels to 32×32 pixels to fit the LeNet-5 architecture. In this process, there can be a loss of data in the images which resulted in much lower classification accuracy.

No	Dataset	classes	# training images	# test images	Image size	Accuracy (%)
1	COVID-CT	2	558	140	20×20 pixels	67.14
2	Four_classes	4	960	240	20×20 pixels	41.25
3	Kvasir	8	3200	800	20×20 pixels	30.75

Table 2.1: *Medical Images*

The second biomedical dataset is the "Four_classes", also taken from the CoroNet[45]. In this dataset the chest X-rays were separated into the classes *COVID*, *normal*, *pneumonia bacterial* and *pneumonia viral*. Table 2.1 shows that in total there is 1200 images for training and testing, where each subject has 300 images.

Figure 2.3 shows the original image and also the cropped topmost left corner of the

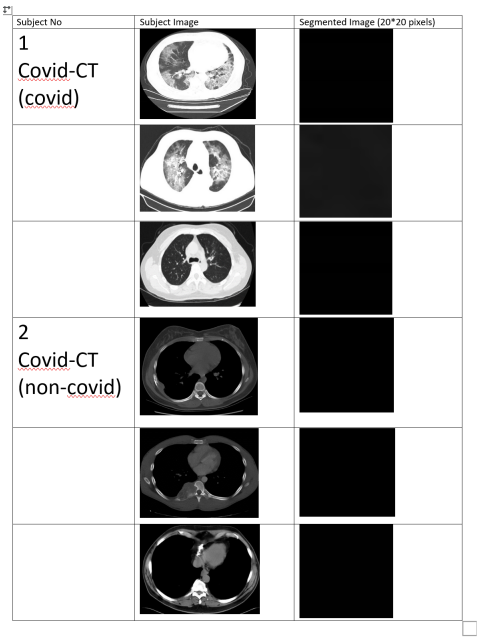

Figure 2.1: *Example images from COVID-CT and a 20×20 portion of the top left corner separated from the original images. Only the sub images were used for the classification.*

original image. As the figure shows, the cropped images are extremely similar and cannot be classified easily by the naked eye. Therefore a convolution neural network was used to distinguish the classes based on only the crop images. The classification accuracy of the group of cropped images is ∼41.25%, which is higher than the random classification of the images that would be ∼25%.

The same architecture when used on the original dataset whose size was reduced to 32×32 pixels to fit the CNN gave a classification accuracy of ∼77.50% whereas the accuracy of classification using the cropped images is ∼41.25%. This shows even though the classification accuracy decreases the CNN can classify some of the images correctly to their label based on only a portion of the background and not the entire image.

The Kvasir dataset[46]is a biomedical dataset that contains images by Endoscopic examinations of the GI tract. The images are collected using endoscopic equipment at Vestre Viken Health Trust (VV) in Norway The data is carefully verified by one or more medical experts from VV and the Cancer Registry of Norway (CRN). The data comprises 4000 images which are divided into 8 classes, each class has 500 images.

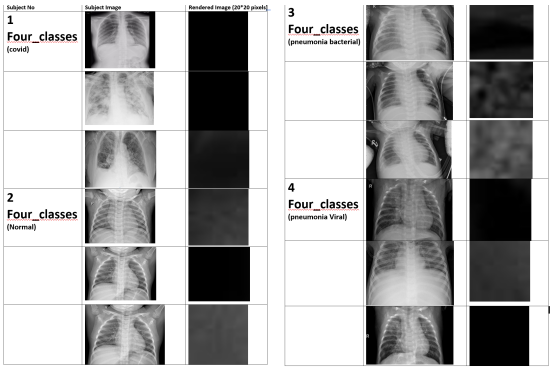

Figure 2.2: *Example images from four_classes and a 20×20 portion of the top left corner separated from the original images. Only the sub images were used for the classification*

Figure 2.2 shows the actual image and the pruned images of size 20×20 pixels that have been cropped from the topmost left corner of the original dataset to classify the images into their respective classes. It is apparent from the Figure 2.2 it is almost impossible to distinctly classify the images into their respective labels without any external help. LeNet-5 architecture apparently can perform this distinction with an accuracy of ∼30.75% even when the images do not give any information about the Endoscopic examinations of the GI tract.

Convolution Neural Network when running on the original dataset whose original size was reduced to 32×32 pixels to fit the model gives a classification accuracy of ∼73.75%. The decrease in the classification is due to the less information provided by only a portion of the background and not the entire image and yet the accuracy of determining the images to the correct labels is greater than randomly assigning the images which show that background also holds a lot of information to classify the images correctly.

2.2.2 Face recognition

For face datasets, the *Yale Faces A* and the *Yale Faces B* were used. The database *Yale Faces A* has 15 subjects where each subject has 11 face images. The *Yale Faces B* has 28 subjects, where each subject has 585 images.

The *Yale Faces B* was transformed into a dataset of the same number of images, where each image in the original dataset was transformed into an image containing the 27×20

Figure 2.3: *Example images from KSVAIR and a 20×20 portion of the top left corner separated from the original images. Only the sub images were used for the classification.*

pixels of the top left corner in the original image. That part of the image contained just the background, which was visually identical in all images. Figure 2.4 and 2.5 shows the five images of the five subjects in each dataset. As the figure shows, the images of all subjects seem identical to the unaided human eye.

In The *Yale Faces A* the background was removed from the image, leading to an artificially blank background. In the case of the Yale A dataset each image was transformed such that the 22×29 pixels from the forehead of each subject were used. Unlike Yale B, in which no pixel containing any feature of the face or hair was used, in Yale A the small images contained pixels representing the skin of the person. However, the images did not contain information that allows identifying the face by visually looking at the image, or to even identify that the image is a face or any other part of a person's body. Figure 2.5 shows examples of the original face images and the smaller images that were used for classification by CNN.

The classification accuracy of the dataset was measured by using the LeNet-5 CNN architecture. The number of epochs is 120. The number of training and test images in each dataset and the classification accuracy of each dataset are shown in Table 2.2.

No	Dataset	classes	# training images	# test images	Image size	Accuracy (%)
1	Yale Faces A	15	132	33	22×29 pixels	54.552
2	Yale Faces B	28	13104	3276	27×20 pixels	87.79

e

Table 2.2: *yaleFaces.*

Although all images are visually similar to each other, CNN was able to classify the

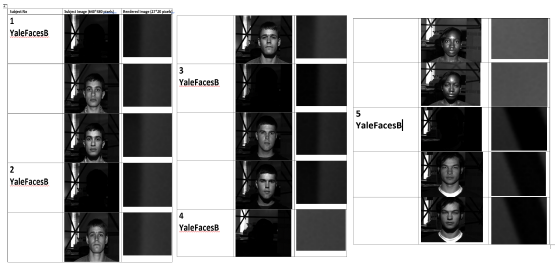

Figure 2.4: *Example images from Yale Faces B and a portion of the topmost left corner separated from the original images. Only the sub images were used for the classification.*

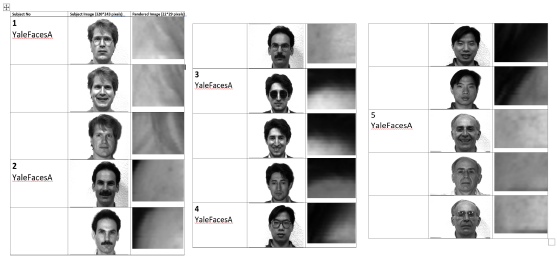

Figure 2.5: *Example images from Yale Faces A and a portion of the forehead images separated from the original images. Only the forehead images were used for the classification.*

images with accuracy far higher than mere chance. With 15 subjects, the mere chance accuracy of *Yale Faces A* is ∼7%, while the mere chance accuracy expected for the *Yale Faces B* dataset is ∼3%. The dramatically higher classification accuracy shows that the CNN identifies discriminating features that are not necessarily related to the faces, and therefore not related to the machine learning problem at hand. That shows that even if the CNN achieves classification accuracy higher than mere chance, it does not necessarily mean that the CNN is indeed able to identify faces, but could identify features of the dataset that allows discrimination between the different subjects.

The CNN algorithm when tested on the original *Yale Faces A* dataset produces a classification accuracy of ∼96.97% which reduces to ∼54.55% when the transformed dataset of segmented images are used. It is prominent from Figure 2.5 that the segmented image is only a portion of the forehead that has the least information about the entire class but the complex convolution neural network can recognize the classes or the object more accurately than a mere chance of classification.

Correspondingly, *Yale Faces B* dataset classifies with an accuracy of ∼99.97% even when

the images are converted to a size of 32×32 pixels. Though there was a significant drop in the accuracy from the original dataset when a cropped images were combined into training sets but the artificial neural network is still able to predict the training dataset by studying the background of the image and not the content descriptor of the original image. This shows the bias in the image dataset as it can be classified even without the real image as shown in Figure 2.4

2.2.3 Object recognition

The object recognition was done on two datasets: *COIL-20* and the *COIL-100*. *COIL-20* contains 20 object classes and each object has 72 images whereas *COIL-100* has 100 subjects and each contains 72 images.[47][48]

A separate subset of each dataset was created from the *COIL-100* and *COIL-20* using the package Image slicer. The image slicer divided each image into 32 parts and only one section of the segmented image was used. The image slice which contains the topmost left side of the original image was used for evaluation purposes. The segmented image has a size of 21×21 pixels whereas the original image of the datasets has a size of 128×128 pixels.

The segmented image contains no information about the object but it is a subsection of the background. The minute underlying difference in each image is impossible to be detected by an unaided eye but complex architecture like LeNet-5 can distinguish those images from each other. The LeNet-5 algorithm uses a complex non-linear function that can differentiate between each target class that is present in the dataset. Figure 2.6 and Figure 2.7 depict the segmented images and the original images alongside each other which is a clear indicator that using a complex convolution neural network makes the task of classification using segmented image possible.

No	Dataset	classes	# training images	# test images	Image size	Accuracy (%)
1	COIL-20	20	1152	288	21×21 pixels	35.42
2	COIL-100	100	5760	1440	21×21 pixels	27.85

Table 2.3: *Datasets used for testing the classification of object recognition benchmarks using deep neural networks.*

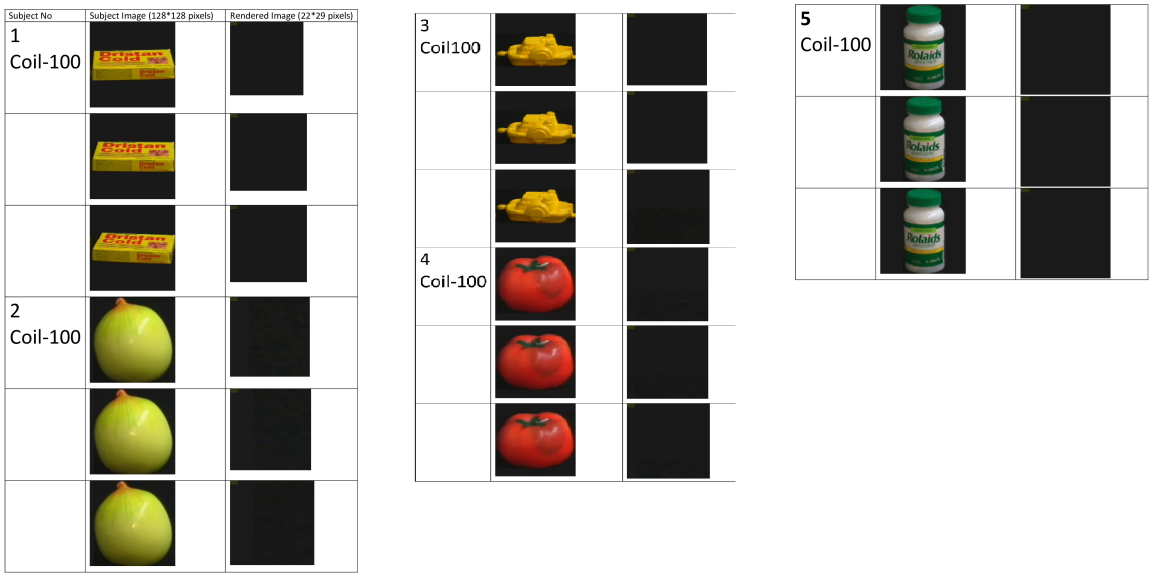

Figure 2.6: *Example images from COIL-100 and the seemingly blank images of the background separated from the original images. Only the blank sub-images were used for the classification.*

The details about the number of images used for training and testing purpose is given in the Table 2.3. For *dataset COIL-20* and dataset *COIL-100 dataset* the accuracy using random choice is ∼5% and ∼1% respectively. The higher rate of accuracy using CNN clearly illustrates that the model can capture the hidden compounded relationship between the segmented image and the target label. CNN with the help of intricate architecture can retrieve information in the pixels of the background image and this knowledge acquired helps to classify the task with higher accuracy.

The classification accuracy of *COIL-20* dataset using un-segmented images comes around ∼98.61%, whereas the classification accuracy when using segmented images drops down to ∼35.42%. Likewise, the classification accuracy of the undivided *COIL-100* dataset is ∼96.46 %, but when the dataset is divided the classification accuracy falls to ∼27.48%. It is clear that there is a substantial decrease in classification accuracy when using the doctored dataset, but it is still able to differentiate between some of the images and not assigning labels using random chance.

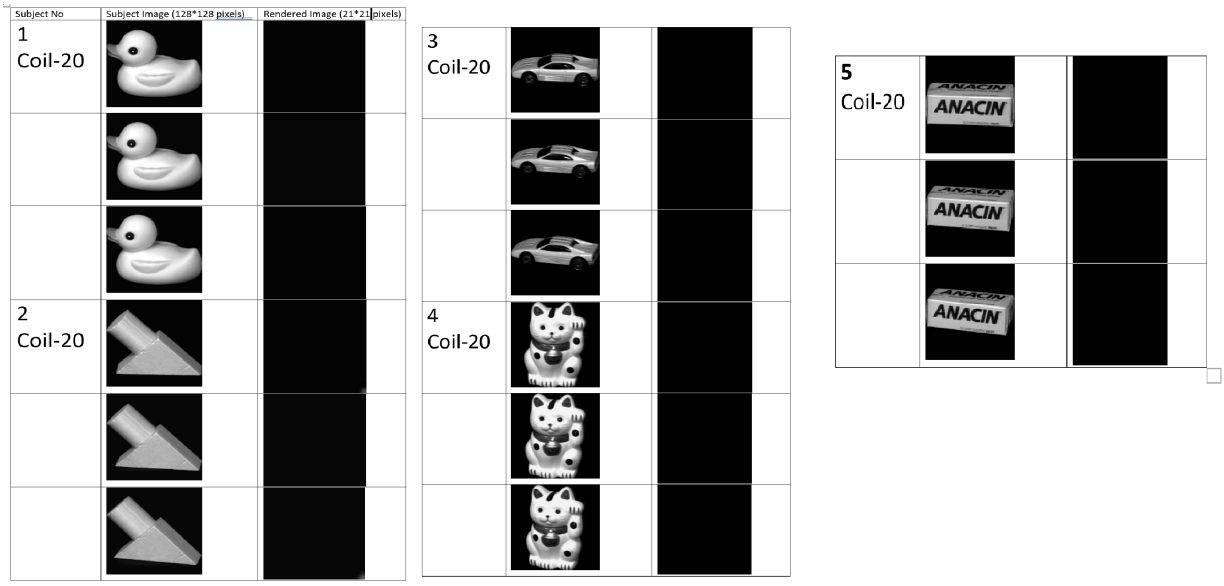

Figure 2.7: *Example images from COIL-20 and the seemingly blank images of the background separated from the original images. Only the blank sub-images were used for the classification.*

2.3 Proposed solutions to CNN classification bias

One of the premier advantages of convolution neural networks (CNNs) is their innate power to select a feature map automatically when supplied with training images. However, the downside of that nature might in some cases lead to potential weaknesses. The automated process of feature map selection without human interference may lead to the use of features that are not necessarily a reflection of the image analysis problem at hand. The classification accuracy provided by the CNN can therefore in some cases be misleading.

Several practices can be used to avoid misleading results due to classification bias driven by consistent yet irrelevant features. Firstly, the background of an image can provide substantial information about the soundness of the image acquisition process. By separating small seemingly blank sub-images of the background we can create a control dataset made with just background information. The ability of a CNN to identify the correct class based on the background alone can alert on the existence of certain anomalies in the data acquisition process. These anomalies are difficult to detect, but CNN can use them to make a classification at accuracy higher than its actual ability to classify these images when anomalies are

not present. That is if a CNN can predict the class of an image based on its background with accuracy higher than mere chance, the overall classification accuracy achieved by that CNN on the entire dataset might be biased, and therefore no strong assumptions can be made on the ability to use that CNN as a valid solution.

Another approach that can be used in acquiring the training set and test set in two separate data acquisition sessions. The common practice of acquiring the entire dataset and then randomly splitting the data into training and test sets can allow CNNs to make a stronger classification accuracy by using information from the imaging session. Separating the acquisition of the training and test sets will ensure that no features that can identify the session in the training set can be used by the CNN to identify the images in the test set. That is if each class of images is acquired in a single imaging session, and then separated into training and test samples, a CNN can associate an image to its session to increase its ability to correctly classify the images. If all test samples are acquired in a different session than the training images, the session information cannot be used to associate test samples with training samples of the same class.

Avoiding the acquisition of data in sessions can also improve the reliability of benchmark datasets used by CNN's. For instance, if each sample is acquired in a separate session, the CNN will not be able to use the subtle but significant information that reflects the imaging session. Imaging each class in a single separate session is a risky practice that can allow CNN's classify the imaging session (e.g., lighting conditions, the temperature of the CCD, etc) rather than the subjects in the images.

Finally, the use of feature engineering, in which the features are pre-designed and known to the user can avoid using automatically generated feature maps that might partially reflect information that is not relevant to the image problem at hand.

2.4 Conclusion

Dataset bias has been discussed in the computer vision literature in the context of the ability of benchmark datasets to reflect the real world and provide a reliable reflection of the

performance of algorithms when applied to real-world problems. In data science, however, dataset bias can lead to false discoveries, especially when the experimentalist analyzes the data without being aware of potential biases.

Here we study biases that are not driven by a human selection of the samples or preferences in the annotation process but driven by the image acquisition process. These biases are very difficult to identify and are not expected as controlled image acquisition is often expected to control also for the possible biases.

The datasets used in this experiment cover a broad range of image classification problems, ranging from medical image-based diagnostics, face recognition, astronomical images, and object recognition datasets. The main commonality between all experiments is that the original datasets were rendered such that each dataset was converted to a new dataset with just the top left corner. The new datasets contained a seemingly blank background, with no clear visual information about the machine learning problem at hand. However, the CNN was still able to predict with accuracy significantly higher than the expected mere chance accuracy. That shows that the CNN can make use of information that is not relevant to the image classification problem, and mislead the experimentalist to believe that the CNN can identify between the different image classes. The CNN makes at least a partial use of background data irrelevant to the visual content of interest, but due to the nature of CNN, the impact of the background is difficult to identify.

Removing the background from the images will not provide a solution to the problem, or make the datasets unbiased. The presence of a signal in the background merely allows isolating the signal from the foreground area, which changes between classes. But the presence of the signal in the background indicates that the signal might also present in the foreground, making it difficult to know whether the information used by the CNN for prediction is based on the relevant visual differences between the classes, on the subtle biases, or a combination of both.

A simple artificial neural network can make the association between image data and labels even without the presence of the visual content by which the images were labeled. That can be explained by the signal originating from the image acquisition process. For instance, if

all subjects of a certain class were imaged during a single imaging session, while all subjects of another class were imaged during another imaging session, the neural network might identify between the imaging sessions rather than between the types of subjects. Subtle lighting conditions or different temperatures of the CCD at the time of imaging can allow a neural network to identify the imaging session, and use that information to predict the correct label.

CNN's learn directly from the image pixels and do not require a step of tailoring specific numerical visual content descriptors. While that advantage makes CNNs much more general and easier to use than virtually any "shallow learning" algorithm, they largely work as "black boxes". The user is not necessarily always aware of what information is being used by the CNN to make its predictions, and therefore using CNNs also has the risk of learning contextual information or other pieces of information that the CNN can use to make better predictions, although not part of the image classification problem at hand.

Therefore, CNNs should be used with caution, and the results provided by CNN should be analyzed carefully. Datasets analyzed by CNNs are sometimes acquired without having CNN analysis in mind, such as biomedical image datasets. In such cases, the application of CNN to these datasets should be done by also using control experiments, such as classifying just the seemingly blank background to ensure the prediction is random.

When acquiring a new image dataset to classify automatically using CNNs, it is suggested to avoid imaging each class in a single imaging session. Imaging samples from the different classes in random order can be more effective for avoiding CNN classification based on the identification of the imaging session rather than the visual content of interest.

Being easy to use, powerful, and accessible through available open-source libraries, CNNs have been becoming extremely popular, and the default solution to image analysis problems. However, while CNN is superior to previous approaches, they also have the downside of overfitting and uncontrolled learning. When a growing population of people who are not machine learning experts uses CNNs, it is important to inform all users also with the possible weaknesses of CNN and avoid experiments that might seem scientifically sound, but in fact, provide biased or unreliable results.

Bibliography

[1] Vanessa Buhrmester, David Münch, and Michael Arens. Analysis of explainers of black box deep neural networks for computer vision: A survey. *CoRR*, abs/1911.12116, 2019. URL http://arxiv.org/abs/1911.12116.

[2] N I Widiastuti. Convolution neural network for text mining and natural language processing. *IOP Conference Series: Materials Science and Engineering*, 662:052010, nov 2019. doi: 10.1088/1757-899x/662/5/052010. URL https://doi.org/10.1088/1757-899x/662/5/052010.

[3] Lin Gui, Yu Zhou, Ruifeng Xu, Yulan He, and Qin Lu. Learning representations from heterogeneous network for sentiment classification of product reviews. *Knowledge-Based Systems*, 124:34–45, 2017. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2017.02.030. URL https://www.sciencedirect.com/science/article/pii/S0950705117301144.

[4] Zhiqiang Geng, Yanhui Zhang, and Yongming Han. Joint entity and relation extraction model based on rich semantics. *Neurocomputing*, 429:132–140, 2021. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2020.12.037. URL https://www.sciencedirect.com/science/article/pii/S0925231220319378.

[5] Quanshi Zhang, Wenguan Wang, and Song-Chun Zhu. Examining cnn representations with respect to dataset bias. 10 2017.

[6] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. URL http://arxiv.org/abs/1602.04938.

[7] Alex A. Freitas. Comprehensible classification models: A position paper. *SIGKDD Explor. Newsl.*, 15(1):1–10, March 2014. ISSN 1931-0145. doi: 10.1145/2594473.2594475. URL https://doi.org/10.1145/2594473.2594475.

[8] Kieran Jay Edwards and Mohamed Medhat Gaber. Astronomy and big data. *Studies in Big Data. Springer*, 2014.

[9] Donald G York, J Adelman, John E Anderson Jr, Scott F Anderson, James Annis, Neta A Bahcall, JA Bakken, Robert Barkhouser, Steven Bastian, Eileen Berman, et al. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3): 1579, 2000.

[10] Nick Kaiser, William Burgett, Ken Chambers, Larry Denneau, Jim Heasley, Robert Jedicke, Eugene Magnier, Jeff Morgan, Peter Onaka, and John Tonry. The pan-starrs wide-field optical/nir imaging survey. In *Ground-based and Airborne Telescopes III*, volume 7733, page 77330E. International Society for Optics and Photonics, 2010.

[11] Timothy Abbott, Filipe B Abdalla, J Aleksić, S Allam, Adam Amara, D Bacon, Eduardo Balbinot, M Banerji, Keith Bechtol, Aurélien Benoit-Lévy, et al. The dark energy survey: more than dark energy–an overview. *Monthly Notices of the Royal Astronomical Society*, 460(2):1270–1299, 2016.

[12] Roberto E González, Roberto P Munoz, and Cristian A Hernández. Galaxy detection and identification using deep learning and data augmentation. *Astronomy and Computing*, 25:103–109, 2018.

[13] Rehab Ali Ibrahim, Mohamed Abd Elaziz, Ahmed A Ewees, Ibrahim M Selim, and Songfeng Lu. Galaxy images classification using hybrid brain storm optimization with moth flame optimization. *Journal of Astronomical Telescopes, Instruments, and Systems*, 4(3):038001, 2018.

[14] Lior Shamir. Automatic morphological classification of galaxy images. *Monthly Notices of the Royal Astronomical Society*, 399(3):1367–1372, 2009.

[15] Manda Banerji, Ofer Lahav, Chris J Lintott, Filipe B Abdalla, Kevin Schawinski, Steven P Bamford, Dan Andreescu, Phil Murray, M Jordan Raddick, Anze Slosar, et al. Galaxy zoo: reproducing galaxy morphologies via machine learning. *Monthly Notices of the Royal Astronomical Society*, 406(1):342–353, 2010.

[16] Evan Kuminski, Joe George, John Wallin, and Lior Shamir. Combining human and machine learning for morphological analysis of galaxy images. *Publications of the Astronomical Society of the Pacific*, 126(944):959, 2014.

[17] Sander Dieleman, Kyle W Willett, and Joni Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2):1441–1459, 2015.

[18] Lior Shamir. Automatic detection of peculiar galaxies in large datasets of galaxy images. *Journal of Computational Science*, 3(3):181–189, 2012.

[19] Ian Timmis and Lior Shamir. A catalog of automatically detected ring galaxy candidates in panstarss. *The Astrophysical Journal Supplement Series*, 231(1):2, 2017.

[20] C Jacobs, T Collett, K Glazebrook, E Buckley-Geer, HT Diehl, H Lin, C McCarthy, AK Qin, C Odden, M Caso Escudero, et al. An extended catalog of galaxy–galaxy strong gravitational lenses discovered in des using convolutional neural networks. *The astrophysical journal supplement series*, 243(1):17, 2019.

[21] Andrew Davies, Stephen Serjeant, and Jane M Bromley. Using convolutional neural networks to identify gravitational lenses in astronomical images. *Monthly Notices of the Royal Astronomical Society*, 487(4):5263–5271, 2019.

[22] Andrew Schutter and Lior Shamir. Galaxy morphology—an unsupervised machine learning approach. *Astronomy and Computing*, 12:60–66, 2015.

[23] Ignacio Sevilla-Noarbe, Ben Hoyle, MJ Marchã, MT Soumagnac, K Bechtol, A Drlica-Wagner, F Abdalla, J Aleksić, C Avestruz, E Balbinot, et al. Star–galaxy classification

in the dark energy survey y1 data set. *Monthly Notices of the Royal Astronomical Society*, 481(4):5451–5469, 2018.

[24] Ting-Yun Cheng, Christopher J Conselice, Alfonso Aragón-Salamanca, Nan Li, Asa FL Bluck, Will G Hartley, James Annis, David Brooks, Peter Doel, Juan García-Bellido, et al. Optimizing automatic morphological classification of galaxies with machine learning and deep learning using dark energy survey imaging. *Monthly Notices of the Royal Astronomical Society*, 493(3):4209–4228, 2020.

[25] PH Barchi, RR de Carvalho, RR Rosa, RA Sautter, M Soares-Santos, BAD Marques, E Clua, TS Gonçalves, C de Sá-Freitas, and TC Moura. Machine and deep learning applied to galaxy morphology-a comparative study. *Astronomy and Computing*, 30: 100334, 2020.

[26] H Domínguez Sánchez, M Huertas-Company, M Bernardi, D Tuccillo, and JL Fischer. Improving galaxy morphologies for sdss with deep learning. *Monthly Notices of the Royal Astronomical Society*, 476(3):3661–3676, 2018.

[27] Asad Khan, EA Huerta, Sibo Wang, Robert Gruendl, Elise Jennings, and Huihuo Zheng. Deep learning at scale for the construction of galaxy catalogs in the dark energy survey. *Physics Letters B*, 795:248–258, 2019.

[28] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8, 2019.

[29] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020.

[30] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[31] François Chollet et al. Keras. https://github.com/fchollet/keras, 2015.

[32] François Chollet et al. Keras: The python deep learning library. *ascl*, pages ascl–1806, 2018.

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[34] Evan Kuminski and Lior Shamir. A computer-generated visual morphology catalog of 3,000,000 sdss galaxies. *The Astrophysical Journal Supplement Series*, 223(2):20, 2016.

[35] Nicholas Paul, Nicholas Virag, and Lior Shamir. A catalog of photometric redshift and the distribution of broad galaxy morphologies. *Galaxies*, 6(2):64, 2018.

[36] Hunter Goddard and Lior Shamir. A catalog of broad morphology of pan-starrs galaxies based on deep learning. *The Astrophysical Journal Supplement Series*, page In Press, 2020.

[37] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[38] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012.

[39] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. *A Deeper Look at Dataset Bias*.

[40] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[41] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. Data-augmentation for reducing dataset bias in person re-identification. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2015.

[42] Nikita Jaipuria, Xianling Zhang, Rohan Bhasin, Mayar Arafa, Punarjay Chakravarty, Shubham Shrivastava, Sagar Manglani, and Vidya N Murali. Deflating dataset bias using synthetic data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 772–773, 2020.

[43] Lior Shamir, John D Delaney, Nikita Orlov, D Mark Eckley, and Ilya G Goldberg. Pattern recognition software and techniques for biological image analysis. *PLoS Computational Biology*, 6(11):e1000974, 2010.

[44] L Shamir. Assessing the efficacy of low-level image content descriptors for computer-based fluorescence microscopy image analysis. *Journal of microscopy*, 243(3):284–292, 2011.

[45] Asif Iqbal Khan, Junaid Latief Shah, and Mohammad Mudasir Bhat. Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images. *Computer Methods and Programs in Biomedicine*, 196:105581, 2020.

[46] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, pages 164–169, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5002-0. doi: 10.1145/3083187.3083212. URL http://doi.acm.org/10.1145/3083187.3083212.

[47] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-20). *Technical Report CUCS-005-96*, 1996.

[48] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-100). *Technical Report CUCS-005-96*, 1996.