Deep learning for anomaly detection in computer vision

by

Ademola Okerinde

BSc, Obafemi Awolowo University, 2008

MSc, University of Ibadan, 2016

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2022

Abstract

Recently, deep learning (DL) inspired algorithms have performed remarkably well on many tasks such as machine translation, speech recognition and image classification etc. However, existing state-of-the-art algorithms struggle to learn the discriminative signals between normal and abnormal classes in an anomaly setting. This is due to the fact that these signals are subtle which often comes in form of little deviations in color, shape, structure etc. To tackle this problem, I propose a more efficient approach named AdeNet that requires lower computation and storage, making it more practical for use on edge devices. Anomaly detection use cases often involves the problem of class imbalance, a case where there are overwhelming samples (majority class) of one or more classes as compared to the others (minority class). This problem also contributes to the inability of DL algorithms to learn distinguishing signals. To address this, I propose an Encoder-based Generative Adversarial Network (eGAN) that leverages on pre-trained model to learn a separable distribution of these classes. Another associated problem to detecting anomalies is zero-shot learning (ZSL). This occurs as a result of the fact that it is practically infeasible to present to our model all possible instances of anomalies during training. Yet, we want models that are robust to new unseen out-of-distribution (OOD) samples during inference. Here, I employ the concept of contrastive learning (CL) to tackle this problem by using pretext task that learns to push embeddings of dissimilar classes far apart, and pull embeddings of similar classes. This seemingly simple concept forces the network to learn salient visual signals that are generalizable to identifying zero-shot instances.

Keywords: AdeNet, Deep Learning, Anomaly Detection, Class imbalance, Transfer Learning, GAN, nash equilibrium, Contrastive learning, Zero-shot learning, Siamese neural network, Pretext task, Contrastive loss.

Deep learning for anomaly detection in computer vision

by

Ademola Okerinde

BSc, Obafemi Awolowo University, 2008

MSc, University of Ibadan, 2016

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2022

Approved by:

Major Professor
William Hsu

Copyright

© Ademola Okerinde 2022.

Abstract

Recently, deep learning (DL) inspired algorithms have performed remarkably well on many tasks such as machine translation, speech recognition and image classification etc. However, existing state-of-the-art algorithms struggle to learn the discriminative signals between normal and abnormal classes in an anomaly setting. This is due to the fact that these signals are subtle which often comes in form of little deviations in color, shape, structure etc. To tackle this problem, I propose a more efficient approach named AdeNet that requires lower computation and storage, making it more practical for use on edge devices. Anomaly detection use cases often involves the problem of class imbalance, a case where there are overwhelming samples (majority class) of one or more classes as compared to the others (minority class). This problem also contributes to the inability of DL algorithms to learn distinguishing signals. To address this, I propose an Encoder-based Generative Adversarial Network (eGAN) that leverages on pre-trained model to learn a separable distribution of these classes. Another associated problem to detecting anomalies is zero-shot learning (ZSL). This occurs as a result of the fact that it is practically infeasible to present to our model all possible instances of anomalies during training. Yet, we want models that are robust to new unseen out-of-distribution (OOD) samples during inference. Here, I employ the concept of contrastive learning (CL) to tackle this problem by using pretext task that learns to push embeddings of dissimilar classes far apart, and pull embeddings of similar classes. This seemingly simple concept forces the network to learn salient visual signals that are generalizable to identifying zero-shot instances.

Keywords: AdeNet, Deep Learning, Anomaly Detection, Class imbalance, Transfer Learning, GAN, nash equilibrium, Contrastive learning, Zero-shot learning, Siamese neural network, Pretext task, Contrastive loss.

Table of Contents

List of Figures ix

List of Tables x

Acknowledgements xii

Dedication xiii

1 INTRODUCTION 1

1.1 Overall Goal 3

1.2 Novel contributions 3

2 AdeNet: DEEP LEARNING ARCHITECTURE THAT IDENTIFIES DAMAGED ELECTRICAL INSULATORS IN POWER LINES 4

2.1 Introduction 4

2.2 Related Work 5

2.3 Proposed Approach 7

2.3.1 Shallow Learning Approach 7

2.3.2 Deep Learning Approach 8

2.4 Methodology and Experimental Design 10

2.4.1 Dataset 10

2.4.2 Data Pre-processing 10

2.5 Results and Discussion 14

2.5.1 Ablation Studies 16

2.6 CONCLUSION 16

3 eGAN: UNSUPERVISED APPROACH TO CLASS IMBALANCE USING TRANSFER LEARNING . . . 19
3.1 Introduction . . . 19
3.2 Related Work . . . 20
3.3 Methodology and Experimental Design . . . 21
3.3.1 Addressing class imbalance with eGAN . . . 21
3.3.2 Selection of pre-trained image classification weights . . . 22
3.3.3 Dataset . . . 23
3.4 Results and Discussion . . . 24
3.4.1 Results . . . 24
3.4.2 Imbalance ratios . . . 27
3.5 Conclusion . . . 30

4 CONTRASTIVE LEARNING APPROACH TO ADDRESSING ZERO-SHOT LEARNING PROBLEM . . . 31
4.1 Introduction . . . 31
4.2 Related Work . . . 33
4.3 Proposed Approach . . . 34
4.4 Methodology and Experimental Design . . . 35
4.4.1 Dataset . . . 35
4.4.2 Data Preprocessing . . . 35
4.4.3 Experimental results . . . 38
4.5 Results and Discussion . . . 39
4.5.1 Conventional Zero-shot learning . . . 40
4.5.2 Generalized Zero-shot learning . . . 40
4.6 Conclusion . . . 41

5 CONCLUSION . . . 42

5.1 Summary and Review of Contributions 42
5.2 Future Directions 43

Bibliography 45

List of Figures

2.1 AdeNet Deep Learning Architecture 9

2.2 An example of a power line image. 11

2.3 Damaged insulators (top) and transistors (bottom) 11

2.4 Undamaged insulators (top) and transistors (bottom) 12

2.5 Model comparison of the ROC curve on insulator dataset 12

2.6 Grad-CAM heatmap visualizations 13

3.1 Encoder-based Generative Adversarial Network (eGAN) architecture 23

3.2 eGAN's average precision (top left), average recall (top right), average F1-score (bottom left), and average performance (bottom right) on CIFAR-10 test dataset using DenseNet121. 25

3.3 Performance of eGAN on **CIFAR-100** test data using DenseNet121. Only the generator network is pre-trained 26

3.4 Discriminator and Generator loss. 26

4.1 Siamese Neural Network 35

4.2 Pretext Tasks 36

4.3 Pie chart of the various species 37

4.4 Example images of various bee species 37

4.5 Positive pair and Negative pair 40

List of Tables

2.1 Number of samples of damaged and undamaged insulators in the dataset. . . 10

2.2 Confusion matrix for Random Forest on test set 16

2.3 Confusion matrix for MobileNetV2 after 20 epochs for test set 16

2.4 Confusion matrix for MobileNetV2 after 10 epochs for test set 16

2.5 Confusion matrix for AdeNet after 10 epochs for test set 17

2.6 Confusion matrix for AdeNet after 20 epochs for test set 17

2.7 Classification accuracy, F1, precision, recall, and ROC area under the curve using different methods on varying validation set. 17

2.8 Classification accuracy, F1, precision, recall, and ROC area under the curve using different methods on test data (average of 5 folds cross validation). . . 18

3.1 Comparative analysis of different pre-trained models configuration on Generator and Discriminator using CIFAR-100 dataset 23

3.2 CIFAR-10 and CIFAR-100 dataset overview - CIFAR-100 in parenthesis . . . 24

3.3 Confusion matrix of imbalance ratio 1:2500 27

3.4 Confusion matrix of imbalance ratio 1:1000 28

3.5 Maximum precision, recall and F1-score on CIFAR-10 and CIFAR-100 (avg. 5-fold) - CIFAR-100 in parenthesis 28

3.6 Precision, recall and F1-score on pneumonia subset of CheXpert dataset . . . 29

4.1 Bumblebee (B.) Species. (**Zero and few shots are in bold**) 38

4.2 ResNet-101 confusion matrix at 0.5 threshold on testset-ALL (45 species) . . 38

4.3 ResNet-101 confusion matrix at 0.5 threshold on testset (24 species) 39

4.4 ResNet-101 Confusion matrix at 0.5 threshold on 21 unknown test set species - zero-shot learning 39
4.5 F1-scores of zero-shot species pair 39
4.6 ResNet-101 confusion matrix at 0.5 threshold on validation 40
4.7 MobileNetV2 confusion matrix at 0.5 threshold on validation 40
4.8 Number of samples in the dataset. 41
4.9 Classification Accuracy, F1, Precision, Recall when using different methods on validation set. 41
4.10 Classification Accuracy, F1, Precision, Recall on test set using ResNet-101 as feature extractor. 41

Acknowledgments

I would first like to thank my thesis advisor, Dr. William Hsu. Without his patience, support, guide and love, this work would have been more difficult. His consistency in steering me in the right direction is well appreciated. I acknowledge my gratitude to my committee members, Dr. Doina Caragea, Dr. Torben Amtoft and Dr. Susan Brown. Many thanks to Dr. Lior Shamir for notable insights and contributions to my research publications. I would like to thank my beloved family and friends for believing in me and encouraging me throughout this academic pursuit. Finally, thanks to the entire K-State / Manhattan community, I couldn't have done it without you all.

Dedication

I dedicate this PhD work to God Almighty.

Chapter 1

INTRODUCTION

An anomaly is an outcome or value that deviates from what is expected, but the exact criteria for what determines an anomaly can vary from situation to situation. It is very difficult to define what is anomalous, as it varies from one case to another. An intrinsic challenge of anomaly is that they occur very rarely leading to the problem of class imbalance. Many research work have approached this problem by simultaneously exploring both normal and abnormal samples.

This thesis is focused on developing algorithms, techniques and methods for detecting anomalies in image and video data. Anomaly detection in spatial-temporal data has wide ranging applications in many domains. For instance, many manufacturing factories perform quality checks during production. This process can be automated for efficiency, however, existing anomaly detection systems which mostly rely on feature engineering suffer from a lot of challenges, one of which is class imbalance. Also, the feature engineering process often requires domain experts for annotation. In my research, we explore deep learning techniques to address the limitation of traditional machine learning approaches to anomaly detection. In addition, in medical diagnosis, we use generative models to transfer knowledge from domains where data is abundant to other domains where there is imbalance, imperfect and limited samples.

The first part of my work is focused on detecting damaged insulators in power lines.

Ceramic insulators are important to electronic systems, designed and installed to protect humans from the danger of high voltage electric current. However, insulators are not immortal, and natural deterioration can gradually damage them. Therefore, the condition of insulators must be continually monitored, which is normally done using UAVs. UAVs collect many images of insulators, and these images are then analyzed to identify those that are damaged. To identify damaged insulators, I developed AdeNet – a deep neural network designed specifically for that task, and tested several approaches to automatically analyze the condition of insulators. Several deep neural networks were tested, as well as shallow learning methods. While the method cannot fully replace human inspection, its high throughput can reduce the amount of labor required to monitor lines for damaged insulators and provide early warning for replacing damaged insulators.

In the follow-up work, I tackled the class imbalance problem that is intrinsic to many anomaly detection task. Class imbalance is an inherent problem in many machine learning classification tasks. This often leads to learned models that are unusable for any practical purpose. To address that problem, an encoder-based Generative Adversarial Network (eGAN) was proposed, which modifies the generator of a GAN by introducing an encoder module and adopts the GAN loss function to directly classify the majority and minority class. That approach eliminates the epistemic uncertainty in the model predictions, as $P(minority)$ and $P(majority)$ need not sum up to 1. The impact of transfer learning and combinations of different pre-trained image classification models at the generator and the discriminator level is also explored. Best result of 0.69 F1-score was obtained on CIFAR-10 classification task with an enforced imbalance ratio of 1:2500.

In another part of my research I addressed the zero-shot learning (ZSL) problem, which is a common solution applied to detect anomalies that are not common in the data. In recent years, self-supervised learning has had significant success in applications involving computer vision and natural language processing. The type of pretext task is important to this boost in performance. One common pretext task is the measure of similarity and dissimilarity between pairs of images. In this scenario, the two images that make up the negative pair are visibly different to humans. However, in entomology, species are nearly indistinguishable and

thus hard to differentiate. In this study, we explored the performance of a Siamese neural network using contrastive loss by learning to push apart embeddings of bumblebee species pairs that are dissimilar, and pull together similar embeddings. Our experimental results show a 61% F1-score on zero-shot instances, a performance showing 11% improvement on samples of classes that share intersections with the training set.

1.1 Overall Goal

My overarching goal is to develop deep learning methods that can address the problem of anomaly detection in computer vision.

1.2 Novel contributions

The AdeNet architecture achieves comparable performance to the state-of-the-art with cheaper computation cost and lower storage demand. This makes it suitable for mobile devices. Epistemic uncertainty is an inherent problem in classification problems due to the fact the predicted probabilities of the classes must sum up to 1. The eGAN approach avoids this, as the outputed score could take up infinite values in the positive or negative space. With the use of pre-trained models and fine-tunning, the cost of computing is lowered and the need to generate synthetic data points, downsample, or oversample are eliminated.

Chapter 2

AdeNet: DEEP LEARNING ARCHITECTURE THAT IDENTIFIES DAMAGED ELECTRICAL INSULATORS IN POWER LINES

2.1 Introduction

Power line insulators deteriorate over time because they are continuously exposed to the weather, including heat, sun, and moisture. Deterioration can eventually damage insulators, so they no longer function. Therefore, identifying damaged insulators is a critical safety task. However, changes in characteristics such as color do not necessarily indicate that the insulator is nonfunctional. Therefore, identifying damaged insulators requires a diverse dataset of different types of damaged and undamaged insulators, at different ages and under different conditions. In this research, we explore various techniques that classify the condition of electrical insulators in power lines.

Automatically detecting anomalies in image data has a broad range of applications[1;2]: automatically identifying product/equipment defects, medical diagnostics[3;4], quality checks at factories, and routine maintenance procedures. Models have been developed that can provide accurate image data. While most deep learning frameworks classify images well, detecting outliers and anomalies is more challenging. Training data for such tasks is often unbalanced because the number of anomaly samples, in many cases, is less than the number of regular samples. Additionally, in some cases, the distinguishing characteristics of regular samples and anomalies are subtle; cracks, flashovers, scratches, and the like are difficult to identify. An effective anomaly detection model must, therefore, learn to identify these details as distinguishing characteristics.

We developed a deep learning architecture, named AdeNet, to identify damaged power line insulators without pre-training. AdeNet also eliminates data augmentation overhead as discussed in[5]. As a result, our model requires less computation and storage, making it more suitable for the computing capabilities of mobile/embedded devices such as unmanned aerial vehicles (UAV). Analysis on the device can be used to decide to acquire additional data while the UAV is still on its mission. AdeNet compares favorably with state-of-the-art deep learning (DL) architectures and shallow learning (SL) techniques. Our experimental results also show that the ROC score is not the best performance metric for an unbalanced dataset. A harmonic mean of recall and precision as expressed by the F1 score is a better choice.

The rest of the paper is laid out as follows. In section 2, we describe the related work. Our proposed approach is discussed in section 3 while detailed experimental setup with dataset used as well as the results are explained in sections 4 and 5. We conclude in section 6 with summary of our findings and recommendations.

2.2 Related Work

Detecting anomalies or outliers in image data is important to computer vision. This task is often characterized by a number of challenges, so it differs from typical supervised machine

learning. One of these challenges is that anomaly samples are far fewer than regular samples.

Many methods have been proposed to address this challenge at both the data and algorithm levels. In[6], a comprehensive real-world dataset is curated for unsupervised detection of anomalies. A common approach is to train the classifier solely on the normal samples with the hope that the architecture will be robust enough to capture the intrinsic characteristic features of the normal class and thus can identify abnormal characteristics by inference[7]. Notable state-of-the-art methods include AnoGAN[8;9], L2 and SSIM Autoencoder[6;10], CNN Feature Dictionary[11], GMM-Based Texture Inspection Model[12], and Variation Autoencoder.

Because untreated damaged insulators are inherently dangerous, automatic monitoring of power line insulators have been substantially researched[13;14]. A basic approach for detecting insulators in aerial images used morphology analysis with Otsu threshold followed by a support vector machine (SVM) classifier[15]. Another approach detected edges and corners represented through MultiScale MultiFeature descriptors[16]. Lattice detection has also been tested with good results in detecting anomalies in insulators in power lines[17].

In the specific context of automatically detecting damaged insulators, Structural Similarity[18] can determine the degree of distortion in damaged insulators from undamaged ones. However, as shown in our experiment, this approach works well if the distortion is digital or artificial, not the result of natural degradation. Circular GLOH-Like descriptors were used for detecting and classifying insulators[19]. A simple method based on shape and the distribution of brightness was proposed as a way to automatically identify cables damaged by lightning[20]. Common deep neural network architectures such as SSD[21] and ResNet[22] have also been used to identify damaged power lines in aerial images[13].

[19] proposed an effective and reliable method of detecting anomalies in insulators using a deep learning technique for aerial images. In the proposed deep detection approach, the single shot multibox detector (SSD), a powerful deep meta-architecture, using two-stage fine-tuning could automatically extract multi-level features from aerial images, thus obviating manual extraction. Inspired by transfer learning, a two-stage fine-tuning strategy was implemented with separate training sets. In the first stage, the basic insulator model was obtained by fine-tuning the COCO model with aerial images, including different types of insulators and

various backgrounds. In the second stage, the basic model was fine-tuned using training sets with specific insulator types and specific situations. After the two-stage fine-tuning, the well-trained SSD model can directly and accurately identify the insulator in aerial images.

According to[23], the insulator is important to a transmission line, and detecting defects in insulators relies heavily on insulator position. Traditional methods of insulator recognition depend on color features and geometric features. These methods are influenced by illumination and background, among other things, resulting in poor generalization.[23] proposed a method where insulators were recognized using deep learning algorithms. First,[23] constructed a training dataset with three categories: insulator, background, and tower. Second,[23] initialized the convolution neural networks as a six-level network and adjusted training parameters to train the model. Finally, the trained model was used to predict the candidate insulator position. With the help of a non-maximum suppression algorithm and the line fitting method,[23] identified the exact location of an insulator.

2.3 Proposed Approach

Paradigms of automatic image classification can be broadly divided into DL and SL. DL often requires more difficult training and a large training set, but it has the advantage of being non-parametric. It is often more accurate than SL. SL requires fewer training samples and is normally less prone to overfitting. However, image features from each image can be computationally expensive, and the resulting machine learning model is not always as accurate as models based on deep learning. In this research, we attempted both DL and SL to compare the two approaches in automatically detecting damaged power line insulators.

2.3.1 Shallow Learning Approach

For shallow learning, we tested the Udat shallow learning image analysis tool[24], which implements the Wndchrm method[25;26]. Udat works by first extracting a large set of 2841 numerical image content descriptors from raw pixels and transforming the raw pixels[27]. The numerical

image content descriptors are shape, edges, textures, polynomial approximation, statistical distribution of pixel values, and others for a comprehensive numerical representation of the image[28]. Detailed information about the Wndchrm algorithm can be found in[25–28].

Once the numerical image content descriptors are computed for all images, the values are then used for mature classification algorithms like Random Forest, Support Vector Machine, Gradient Boosting, and Naive Bayes.

2.3.2 Deep Learning Approach

In deep learning, multiple layers of mini-algorithms, called neurons, work together to draw complex conclusions. We chose the AdeNet architecture and compared it with several deep neural network architectures, from the basic LeNet-5 to more complex ResNet-101, VGG19, and MobileNetV2.

MobileNetV2

MobileNetV2[29] is a mobile architecture based on convolutional neural networks (CNNs). The basic building block is a bottleneck depth-separable convolution with residuals. Our implementation of MobileNetV2, taken from tensorflow-keras applications, has 398, 690 trainable, and 14,000 non-trainable parameters. The total size of the model is 5.7MB.

AdeNet

AdeNet is a deep learning architecture implemented with three layers of CNNs, each with batch normalization, maxpooling, and ReLU with no dropout. One fully connected layer comes before the softmax layer. Figure 2.1 shows the architecture. It contains the initial fully convolutional layer with 32 filters. We always used kernel size 3×3 as is standard for modern networks. The trained model size was 1.3MB with 102,082 trainable parameters and 448 non-trainable parameters.

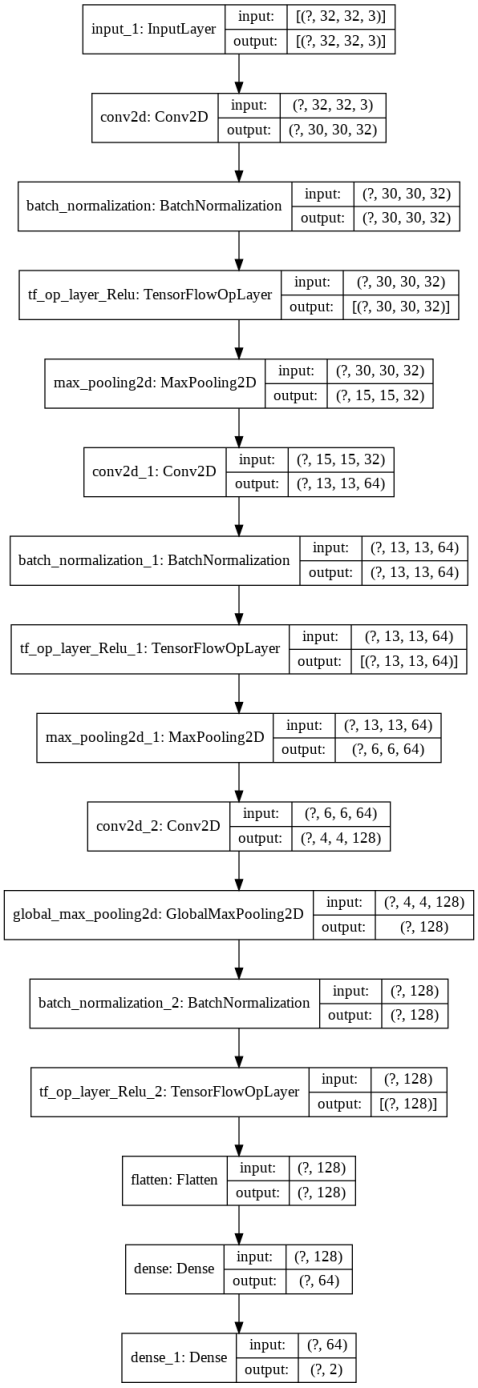


Figure 2.1: *AdeNet Deep Learning Architecture*

2.4 Methodology and Experimental Design

2.4.1 Dataset

The dataset was provided by Black and Veatch, consisting of 1696 JPEG images of power lines with the dimensions of 5280×3956 and a resolution of 72x72. We randomly divided the entire dataset into 80% training sets and 20% test sets. Each image is annotated with a bounding-box around the objects of interest: the insulators. The insulators are broadly categorized as damaged (positive class) and undamaged (negative class). Table 4.8 shows the number of samples in each class after cropping out the insulators.

Table 2.1: *Number of samples of damaged and undamaged insulators in the dataset.*

Dataset	Damaged	Undamaged	#images
Train + Validation	1417	2836	1484
Test	290	835	212

The dataset presents a number of interesting challenges: within class variations, class imbalance, noisy background, and varying orientations and scales. Using within class variations as an example, in our experiments, we classified as damaged any insulator with flash-over or that was broken, missing, or fried. This leads to skewness of samples within each class as well.

2.4.2 Data Pre-processing

Figure 2.2 shows an example of a power line image. The electrical insulators were then cropped out of images for feature extraction, giving us a total of 4253 images for training and validation as well as 1125 images for testing. Figure 2.3 shows samples of damaged insulators, and figure 2.4 shows samples of undamaged insulators. For SL, we used Udat to extract the 2900 most informative image features from the insulators. Udat uses statistical techniques like Radon transform features, Chebyshev Statistics, and Multi-scale Histograms for image feature extraction. For DL, because they varied in scale, we padded insulators to the maximum dimension in each batch during training and testing.



Figure 2.2: *An example of a power line image.*

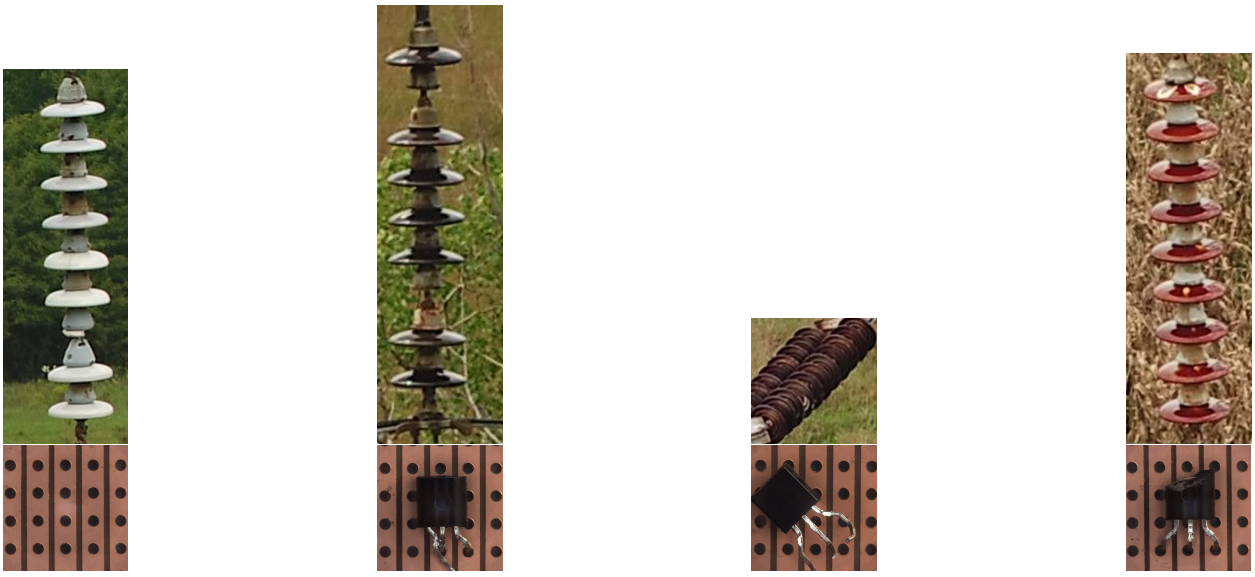


Figure 2.3: *Damaged insulators (top) and transistors (bottom)*

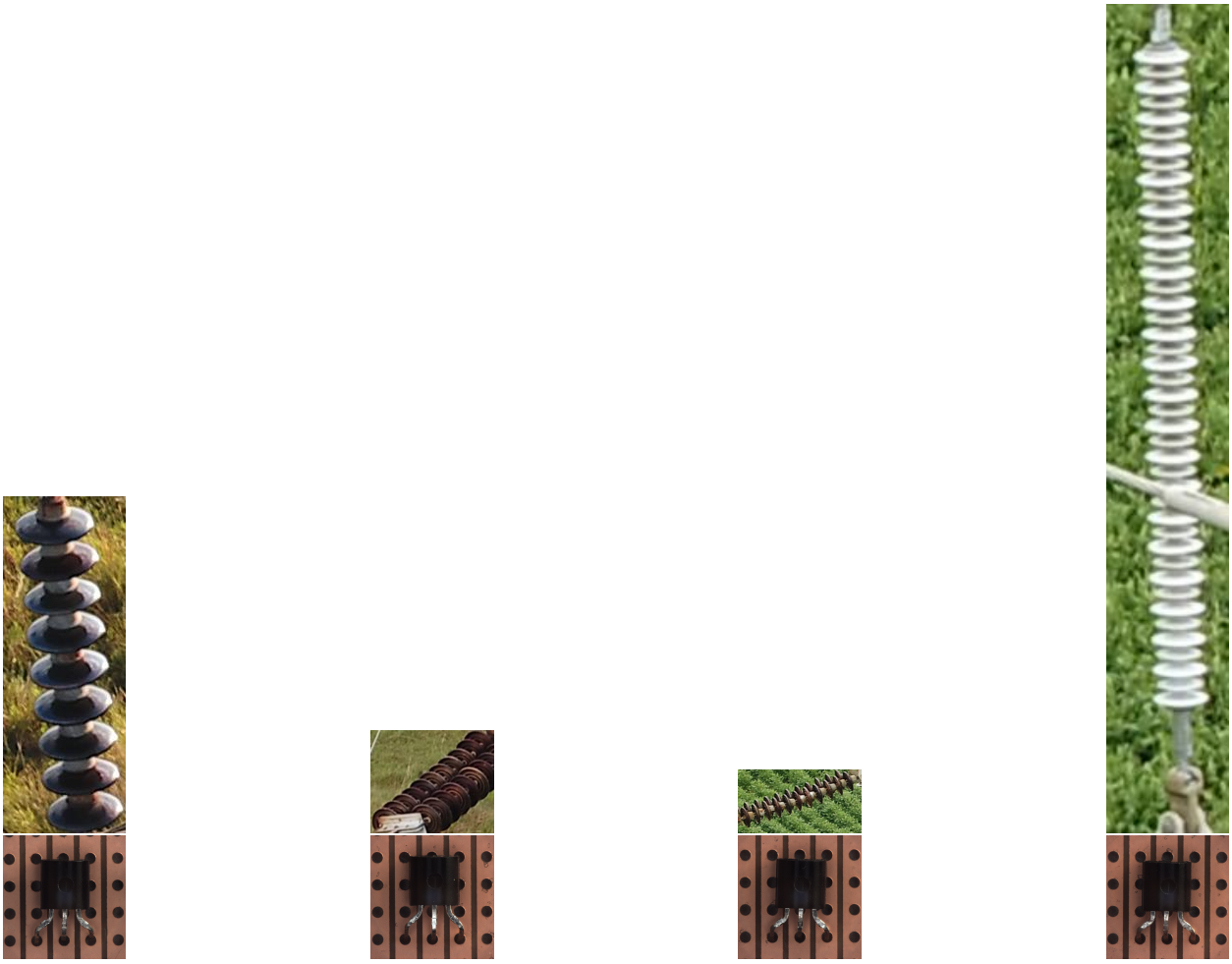


Figure 2.4: *Undamaged insulators (top) and transistors (bottom)*

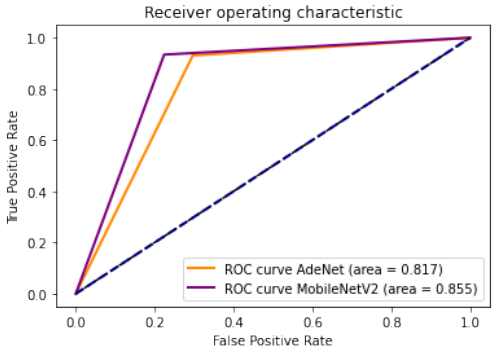


Figure 2.5: *Model comparison of the ROC curve on insulator dataset*

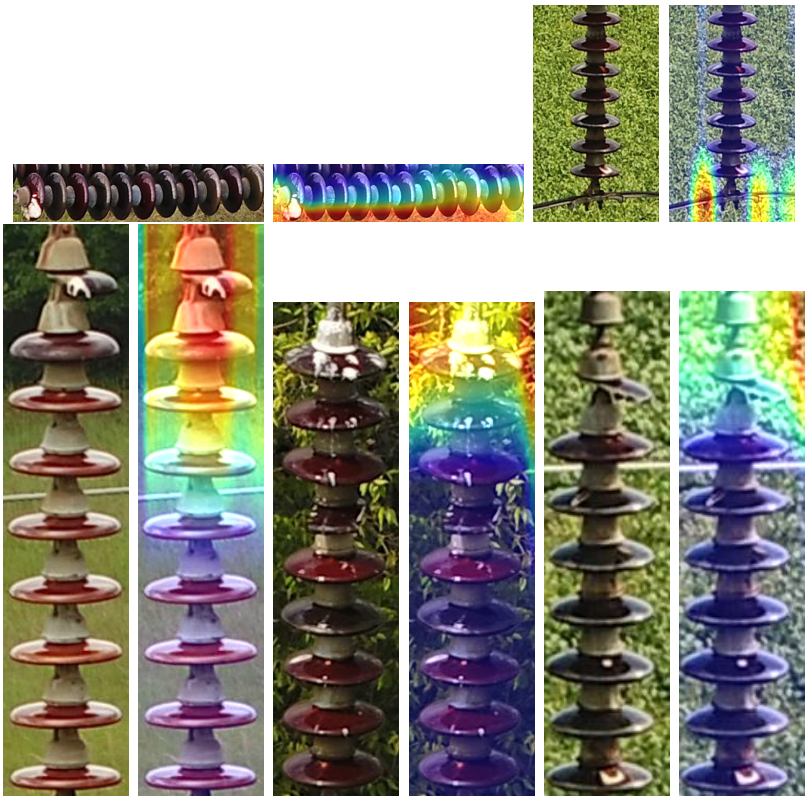


Figure 2.6: *Grad-CAM heatmap visualizations*

2.5 Results and Discussion

The different methods were applied to the dataset described in Section 4.4.1. For training and testing, we used mostly 5-fold as the different classifiers for cross-validation. Each split was trained with 10 or 20 epochs. For transfer learning, we used models pre-trained on the ImageNet dataset. We used a batch size of 16 for training and testing. Most of our DL experiments were run on Tesla P100-PCIE on Google Colab. The DL models were built using TensorFlow[30].

The results are shown in tables 2.7 and 4.9. For the analysis of the results, the tables give accuracy, F1 score, precision, recall, and ROC area under the curve for each classifier. The measurements are the average of all folds.

We found that MobileNetV2 sometimes performed better if trained using more epochs, starting with saved model weights. The performance often dropped if the F1 score for the initial 10 epochs reached 91% in the damaged class. The model often inverted performance from one class to favor another if the results for the first 10 epochs were skewed. As tables 2.3 and 2.4 show, the model had fewer false negatives when trained for fewer epochs. This is crucial for this situation.

Without transfer learning, MobileNetV2 was highly influenced by the class imbalance because most data were classified as negatives. Likewise, LeNet-5[31], EfficientNetB7[32], VGG19[33], and ResNet-101[22] models all classified most data as negatives. All models other than LeNet-5 were pre-trained on the ImageNet dataset[34]. Table 2.7 shows their performance on different validation percentage splits.

Our experimental findings showed that all MobileNetV2 model performances on the test set correlated with results from the validation set. For example, if a model did well on the positive class in validation, it also performed well on the positive class in the test. We ran all five models on the test set, and the results showed that MobileNetV2 performed better on 20-epochs-models on the test set. We reported the average of the five models in Table 4.9.

Comparing AdeNet with similar simple DL architectures like LeNet5 revealed that an overly simple network was not optimal, but a carefully engineered simple architecture that

balanced simple and complex designs, like ResNet-101 and VGG19, worked. The confusion matrices in tables 2.2, 2.3, 2.4, 2.5, and 2.6 reveal that AdeNet was significantly better at classifying both classes of the dataset. This was confirmed by high F1 scores in Table 4.9. F1 score is a better performance metric for evaluating models developed for imbalanced datasets. Consider, for instance, random forest with high ROC score, but poor F1 score (see Table 2.2).

The small number of samples in[6]'s transistor category meant AdeNet performed poorly with and without fine-tuning because the entire dataset was classified as negatives, i.e., undamaged transistors. We borrowed intuition from[35] and experimented with freezing two, ten, and thirteen layers of pre-trained AdeNet on the insulator dataset to see how well it transferred to the transistor detection task. This reinforced the conclusion that a DL model performs better with more data. However, SL with Udat achieved an accuracy of 92.8% (10-fold-cross-validation split) on validation data with roughly half of the damaged class miss-classified.

With Grad-CAM, we used heatmap to visualize the model's attention on the insulator while making predictions as shown in Figure 2.6. Grad-CAM uses the penultimate (pre-dense) convolutional layer containing spatial information that is completely lost in dense layers. The model's reasoning did make some sense from the human visual system perspective. The areas that constituted damage in the insulators were located where the heatmap was most intense, i.e., heavily red.

Figure 2.5 compares the ROC curve for the models using the Black and Veatch insulator dataset. The curve shows that MobileNetV2 was slightly better than AdeNet. This plot was based on one model. Table 4.9 shows that, on average, AdeNet outperformed MobileNetV2.

A major hyper-parameter tuning task was to determine how long to train a deep neural network. We experimented with using early stopping and fixed epochs, finding that performance was much the same. However, with callbacks, training time was significantly reduced.

Of the tested classifiers, the best shallow learning classifier was Random Forest. However, its confusion matrix (see Table 2.2) revealed that the classifier did not do well in classifying

the damaged class, making it the worst classifier for this task.

Table 2.2: *Confusion matrix for Random Forest on test set*

	Predicted Damaged	Predicted Undamaged
Actually Damaged	55	1395
Actually Undamaged	10	4165

Table 2.3: *Confusion matrix for MobileNetV2 after 20 epochs for test set*

	Predicted Damaged	Predicted Undamaged
Actually Damaged	991	459
Actually Undamaged	1063	3112

Table 2.4: *Confusion matrix for MobileNetV2 after 10 epochs for test set*

	Predicted Damaged	Predicted Undamaged
Actually Damaged	724	726
Actually Undamaged	312	3863

2.5.1 Ablation Studies

We inspected parts of the architecture and parameters for analysis.

Batch Normalization. We observed no significant differences in model performance with or without batch normalization[36] possibly because AdeNet is not exceptionally deep. We could thus further reduce the parameter by $\sim$1%.

2.6 CONCLUSION

In this research, we addressed the task of detecting damaged electrical insulators in power line images, an important safety task. Detecting damaged electrical insulators has long been a labor-intensive task requiring substantial experience and careful examination of the images.

The new architecture, AdeNet, based on a deep convolutional neural network has the advantage of requiring little energy, allowing it to be used on low-energy devices like UAVs, which are often used to acquire the data for identifying damaged insulators. The low energy

Table 2.5: *Confusion matrix for AdeNet after 10 epochs for test set*

	Predicted Damaged	Predicted Undamaged
Actually Damaged	947	503
Actually Undamaged	257	3918

Table 2.6: *Confusion matrix for AdeNet after 20 epochs for test set*

	Predicted Damaged	Predicted Undamaged
Actually Damaged	1026	424
Actually Undamaged	213	3962

allows AdeNet to make decision on the UAV, which can then be used to acquire more data of suspect insulators during a mission, without going back to base and re-deploying the UAV. Like intuition in[37;38], we found an architecture that is not as simple as LeNet5 and not as complex as ResNet (just the right mix of simplicity and complexity) that also performs well in detecting anomalies in damaged insulators in power lines.

In comparing the proposed method and other solutions, including shallow learning, we found that deep learning architectures outperform shallow learning architectures. Within DL architecture, fine-tuning the pre-trained CNN models did not improve the performance in all cases. We demonstrated this through experiments with our AdeNet architecture. Experimental results also showed that our architecture, without pre-training, outperformed all other deep neural networks. This makes AdeNet suitable for mobile/embedded devices that have storage and computation constraints.

Table 2.7: *Classification accuracy, F1, precision, recall, and ROC area under the curve using different methods on varying validation set.*

Dataset	Classifiers	Acc	F1	Precision	Recall	ROC Area	#folds
val	Random Forest	0.86	0.85	0.86	0.86	0.92	10
val	MultiLayer Perceptron	0.82	0.82	0.82	0.82	0.86	10
val	Gradient Boosting	0.87	-	-	-	-	
val	LeNet-5	0.68	0.46	0.70	0.53	0.53	5
val	AdeNet + 10 epochs	0.86	0.83	0.87	0.81	0.81	5
val	EfficientNetB7	0.63	0.36	0.36	0.49	0.49	5
val	VGG19	0.67	0.40	0.44	0.50	0.50	5
val	ResNet-101	0.59	0.40	0.35	0.51	0.51	5

Table 2.8: *Classification accuracy, F1, precision, recall, and ROC area under the curve using different methods on test data (average of 5 folds cross validation).*

Learning	Classifiers	Acc	F1	Precision	Recall	ROC Area
Shallow	Random Forest	0.79	0.64	0.79	0.62	0.82
Shallow	Random Tree	0.69	0.60	0.60	0.60	0.60
Shallow	Naive Bayes	0.56	0.56	0.64	0.67	0.76
Shallow	MultiLayer Perceptron	0.77	0.69	0.69	0.68	0.79
Shallow	Support Vector Machine	0.72	0.65	0.65	0.66	0.66
Deep	MobileNetV2 + 10 epochs	0.82	0.72	0.77	0.71	0.73
Deep	MobileNetV2 + 20 epochs	0.70	0.63	0.76	0.70	0.71
Deep	AdeNet + 10 epochs	0.86	0.81	0.85	0.80	0.80
Deep	**AdeNet + 20 epochs**	**0.89**	**0.84**	**0.87**	**0.83**	**0.83**

Acknowledgement

We are grateful to Kansas State University Beocat team for providing us with access to the high performance computing cluster for running the experiments.

Chapter 3

eGAN: UNSUPERVISED APPROACH TO CLASS IMBALANCE USING TRANSFER LEARNING

3.1 Introduction

A dataset is considered imbalanced when there is a significant, or in some cases, extreme disproportion between the number of samples of the different classes in the dataset. The class or classes with large number of samples are called the majority, while the class with few examples are denoted as the minority. In many cases, the machine learning model is required to correctly classify the minority class while minimizing the misclassification of the majority class. However, the skewness in the data often leads machine learning classification methods to favour the majority class.

Class imbalance problem in computer vision is normally approached either at the data level or algorithm level. Using data augmentation, a class with a small number of samples can be expanded into a class with a much larger number of samples. Earlier data augmen-

tation was achieved simply by transforming images via scaling, cropping, flipping, padding, rotation, brightness, contrast, saturation level etc[39]. Now-a-days, synthetic images can also be generated using generative models such as VAE, GAN[40 41]. As a result, a humongous image dataset can be created from the images of the minority class.

At the algorithm level, the objective function is tweaked to heavily penalize the network for mis-classifying the minority class[42 43]. The most popular is cost-sensitive approach. Here, the classifier is modified to incorporate varying penalty for each of considered groups of examples. By assigning a higher cost to the less represented set of samples its importance is boosted during training.

Transfer learning has been known to help improve the performance of machine learning models[44]. By fine-tuning varying number of layers in the pre-trained image classification model, the pre-trained model can serve as a feature extractor, while adding a classifier head for more specific feature learning for the current task.

In this work, we compared the performance of various pre-trained image classification models for the task of unsupervised image classification with varying imbalance ratios. Our architecture, named eGAN, is developed to serve as a basis for this comparison. Using GAN[41], we reparameterise the task of the discriminator as a classifier which outputs a positive score for majority samples and a negative score for the minority ones. We integrate an encoder module to the GAN network that encodes the minority samples into a latent code from which the generator learns. While most GAN-based architectures focus on the output of the generator, in our proposed approach, as we intuitively adapted the vanilla GAN network and the corresponding loss function to directly classify the majority and the minority class, we are more concerned about the performance of the discriminator.

3.2 Related Work

There have been a lot of work in the last few decades to address class imbalance. Earlier approaches include deliberate undersampling of the majority class or oversampling of the minority class by mere copying[45]. However, for image data the earlier approach leads to loss

of useful data information while the latter approach causes overfitting[46]. Data augmentation via rotation, scaling, cropping etc can be considered as a variant of oversampling which copy the same data, however with little modification[39]. VAE and GAN enables the generation of completely new data[40 41]. In recent years, GAN-based approaches have gained much popularity than others and a good number of variants of vanilla GAN have been proposed to address class imbalance[47 48 49].

In[50], an ensemble method was proposed based on advanced generative adversarial network to generate new samples for the minority class to restore balance. Our opinion is that the computational demand of such approach is enormous, and many low-income countries of the world do not have access to such computation power. Deep cascading (DC) with a long sequence of decision trees could help to handle unbalanced data[51]. A DC is a sequence of n classifiers where each sample x passes to the next classifier only if the current one classifies it as positive according to a high-sensitivity decision threshold. However, this works well with foreground-background imbalance unlike the classification task. Transfer learning with GAN was used to generate images from limited data in[44]. Their result showed that knowledge from pre-trained networks can ensure faster convergence and significantly improve the quality of generated images.

3.3 Methodology and Experimental Design

In this section, we discuss our proposed approach and the various testbeds that were used in our experiments.

3.3.1 Addressing class imbalance with eGAN

The proposed architecture is based on adaptation of existing Deep Convolutional Generative Adversarial Network (DCGAN)[47] by incorporating an encoder module. This module encodes minority samples in latent space needed by the generator G to generate minority samples that are capable of fooling the discriminator D. On the other hand, the discriminator D

is fed with data samples drawn from majority distribution and the generated output of the generator G. D and G are simultaneously optimized through the following two-player minimax game $\overset{min}{G}\ \overset{max}{D}\ V(D,G)$ with the value function V(D, G) in 3.1.

$$V(D,G) = E_{X_{ma}\sim P_{ma}}[\log D(X_{ma})] + E_{X_{mi}\sim P_{mi}}[\log(1 - D(G(X_{mi})))] \tag{3.1}$$

where X_{ma} and X_{mi} are majority and minority sample distributions respectively.

Over the course of iteration, the discriminator D is optimized to assign a negative score to the minority data distribution and a positive score to the majority data distribution. This enables the discriminator D to act as a classifier.

Encoder-Generator module

Our latent space is composed of 128 units vector. Rather than feeding the generator with random noise as is typical of most GAN implementation, we added an encoder module that forces the generator to learn from known distribution (minority distribution). The encoder part consists of the pre-trained DenseNet121 followed by global average pooling layer and latent dimension space. The generator part has two transposed convolutional layers. We use LeakyRelu activation function with alpha set to 0.2; batch normalization and Sigmoid function at the final layer.

Discriminator module

The pre-trained discriminator has 7,038,529 parameters out of which only 39,937 are trainable. A layer of global average pooling follows the pre-trained DenseNet121. We use a dropout of 0.2 followed by the final one unit dense layer. The overall architecture of our encoder-based generative adversarial network is shown in Figure 3.1.

3.3.2 Selection of pre-trained image classification weights

We perform experiments on VGG16, VGG19, EfficientNetB2, ResNet101 and DenseNet121 pre-trained classification models on ImageNet dataset. Here, we fine-tuned only top five

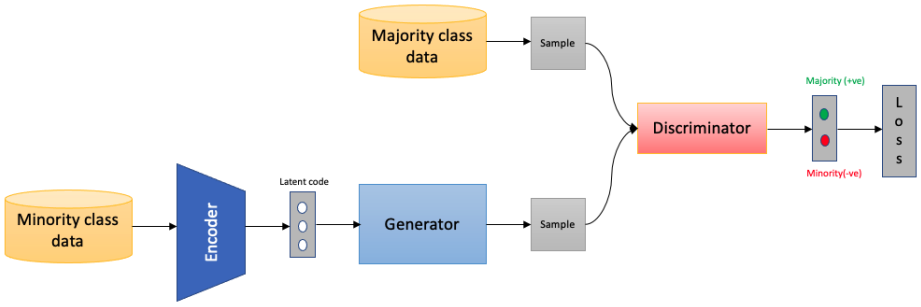


Figure 3.1: *Encoder-based Generative Adversarial Network (eGAN) architecture*

layers at each of the pre-trained models. Table 3.1 shows the maximum precision, recall and F1-score obtained on CIFAR-100 with imbalance ratio 1:50 by using different combinations of pre-trained models.

Table 3.1: *Comparative analysis of different pre-trained models configuration on Generator and Discriminator using CIFAR-100 dataset*

Discriminator pre-trained	Generator pre-trained	Precision	Recall	F1
ResNet101	VGG19	0.72	1.0	0.78
VGG19	ResNet101	0.73	1.0	0.69
EfficientNetB2	VGG19	1.0	0.22	0.32
VGG19	EfficientNetB2	0.7	0.86	0.71
ResNet101	VGG16	1.0	0.17	0.27

DenseNet121[52] was used for pretraining our eGAN. After experimenting different pre-trained architectures and different layers of fine-tuning, we obtained best result with fine-tuning only top 5-layer out of 427 layers of DenseNet121.

3.3.3 Dataset

Several commonly used datasets were used in this study. In order to model the real-world scenario of heavy imbalance, we used only few samples of the minority class as input to the encoder module. Detail overview is shown in Table 3.2.

The CIFAR-10 dataset[53] consists of 60,000 32×32 colour images in 10 classes, with 6000 images per class. There are 50,000 training images and 10,000 test images. Here we use

Table 3.2: *CIFAR-10 and CIFAR-100 dataset overview - CIFAR-100 in parenthesis*

Class imbalance ratio	# training minority class	# training majority class	# testing minority class	# testing majority class
1:2500	2(-)	5000(-)	1000(-)	1000(-)
1:1000	5(-)	5000(-)	1000(-)	1000(-)
1:500	10(1)	5000(500)	1000(100)	1000(100)
1:50	100(10)	5000(500)	1000(100)	1000(100)
1:1	500(500)	500(500)	1000(100)	1000(100)
*1:1	-(1)	-(1)	-(100)	-(100)

airplane as minority and *automobile* as majority.

CIFAR-100[53] is similar to CIFAR-10, except it has 100 classes containing 600 images each. Each class has 500 training images and 100 testing images. The 100 classes in the CIFAR-100 are grouped into 20 superclasses. Each image comes with a "fine" label (the class in which it belongs) and a "coarse" label (the superclass). We also use the pneumonia subset of Stanford CheXpert dataset[54] for experimenting on an inherent imbalanced dataset. The dataset contains 4576 and 167407 minority and majority samples, respectively.

3.4 Results and Discussion

3.4.1 Results

All discriminator scores that are less than zero are classified as minority, otherwise they are classified as majority class. Table 3.1 shows the result obtained by combining various pre-trained models. Adam optimizer with a learning rate of 1e-4 was used to train all models for 100 epochs.

As can be seen in Figure 3.2, the model achieved a Nash equilibrium on test data at around 10 epochs. Here, we perform inference on test data at every epoch and plot the number of samples correctly classified. Five layers of DenseNet121 were fine-tuned at each generator and discriminator module, while 422 layers' weight were kept fixed. At Nash, the discriminator correctly classified roughly 700/1000 of each of the minority and majority test

data. This result convinces us that transfer learning with GAN can be used to overcome the challenge of highly imbalanced dataset, owing to the fact that we train only with 10 samples of the minority class and 5000 samples of the majority class. Similar performance is observed in CIFAR-100 with imbalance ratio 1:50.

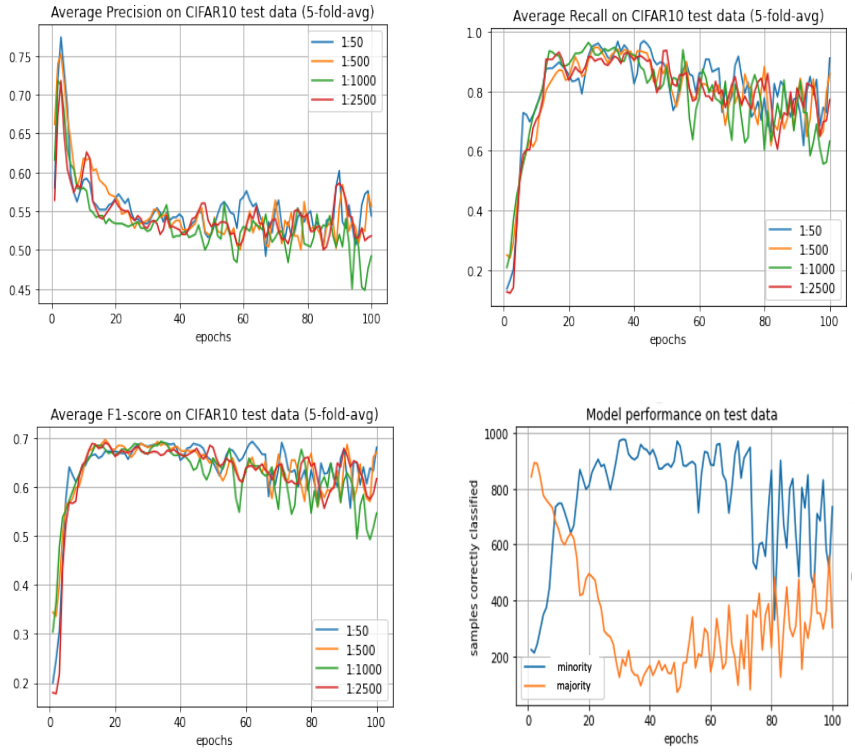


Figure 3.2: *eGAN's average precision (top left), average recall (top right), average F1-score (bottom left), and average performance (bottom right) on CIFAR-10 test dataset using DenseNet121.*

Without pre-training the discriminator, the effect of the high imbalance in the training set is revealed, as the discriminator is skewed towards the majority class in the training set, thereby missing all the minority samples in the test data. This can be seen in Figure 3.3 on CIFAR-100. This behaviour pattern is observed on CIFAR-10 as well. We experiment with no pre-training at all, neither in the discriminator nor generator, and observed exact same pattern. Therefore, we can safely conclude that the use of transfer learning helps unsupervised image classification in a highly imbalanced domain.

Training can be stopped as soon as Nash equilibrium is reached, as this point gives the model best performance on the minority and majority class. An acceptable threshold can

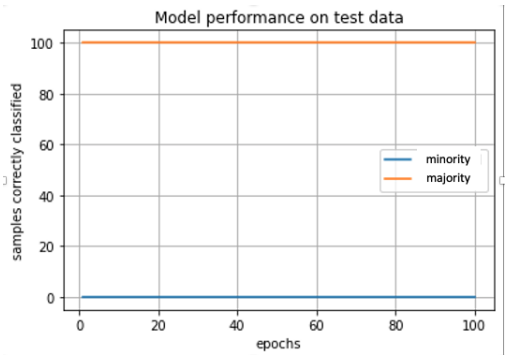


Figure 3.3: *Performance of eGAN on **CIFAR-100** test data using DenseNet121. Only the generator network is pre-trained*

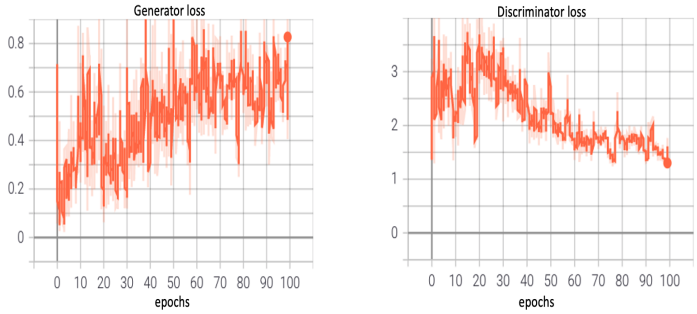


Figure 3.4: *Discriminator and Generator loss.*

also be set for the absolute difference of the number of correctly classified samples of both classes. For instance, if the | correctly_classified_minority - correctly_classified_majority | $\leq$ 20. The precision, recall and F1-score curves on CIFAR-10 averaged over five folds at different imbalance ratios are shown in Figure 3.2.

We observed that at the early training epochs, typically between 1 and 40 epochs, the generator tries to achieve its objective of fooling the discriminator by generating samples from the majority class fed into the discriminator. That results in more of the minority samples being mis-classified as the discriminator "knows" the distribution of the majority too well. A drastic change occurs when the generator start generating samples from latent vector, which can fool the discriminator as seen in generator and discriminator loss shown in Figure 3.4.

3.4.2 Imbalance ratios

To eliminate bias in model performance, we conducted 5-fold-cross-validation on the minority samples and average the result.

Class ratio of 1:2500

We experiment on CIFAR-10 dataset by deliberately using an unbalanced subset of the training set. At the 4th epoch, our model correctly identifies 257 minority and 821 majority samples out of 1000 each. At epoch 5, a sharp change occurred that led to 821 minority samples being correctly classified, while only correctly classifying 233 majority samples as shown in Table 4.7. We also observed that at epoch 72 the performance of the network on the majority and minority classes reached a nash equilibrium with a threshold difference of less than or equal to 20.

Class ratio of 1:1000

At epoch 81 on CIFAR-10 dataset, nash equilibrium was reached. At this epoch, 532 and 525 minority and majority test data respectively were correctly classified. We observed that the classifier had another major shift between epoch 5 and 6. At epoch 5, best result was obtained. The network was able to classify 863 majority tests and 585 minority tests correctly out of the 1000 samples. At epoch 6, 634 majority and 810 minority tests were classified correctly as shown in Table 3.4. Maximum precision, F1-score and recall of 0.88, 0.74 and 0.99 were obtained at epochs 3, 6 and 37, respectively.

Table 3.3: *Confusion matrix of imbalance ratio 1:2500*

	Predicted minority	Predicted majority
Actual minority	821	179
Actual majority	767	233

Table 3.4: *Confusion matrix of imbalance ratio 1:1000*

	Predicted minority	Predicted majority
Actual minority	810	190
Actual majority	366	634

Class ratio of 1:500

Both CIFAR-10 and CIFAR-100 were used to experiment imbalance ratio 1:500. On CIFAR-10, maximum precision, recall and F1-score on averaging 5-fold-cross-validation are 0.75, 0.95 and 0.70 respectively as shown in Table 3.5. The maximum precision is slightly lower on CIFAR-100 with 0.60. However, the recall and F1-score which are 0.96 and 0.68 are roughly the same.

Class ratio of 1:50

We demonstrate our model performance on imbalance ratio 1:50 using CIFAR-100 and CIFAR-10. For CIFAR-100, a sudden change occurred between epoch 69 and 70 as follows majority: 57, minority: 59; and majority: 55, minority: 59. A Nash equilibrium is attained at epoch 68, with 56 correctly classified minority as well as majority class. At epoch 97 maximum F1-score and recall of 0.66 and 0.77 were obtained respectively, while maximum precision of 0.6 was obtained at epoch 74.

Table 3.5: *Maximum precision, recall and F1-score on CIFAR-10 and CIFAR-100 (avg. 5-fold) - CIFAR-100 in parenthesis*

ratio	Precision	Recall	F1
1:2500	0.72(-)	0.94(-)	0.69(-)
1:1000	0.72(-)	0.96(-)	0.69(-)
1:500	0.75(0.60)	0.95(0.96)	0.70(0.68)
1:50	0.78(0.7)	0.97(0.86)	0.69(0.71)
1:1	**0.82**(0.72)	0.98(**1.0**)	0.72(**0.78**)
*1:1	-(0.53)	-(0.86)	-(0.63)

Table 3.6: *Precision, recall and F1-score on pneumonia subset of CheXpert dataset*

	#training		#testing					
Model	minor	major	minor	major	imbalance ratio	Precision	Recall	F1
eGAN	30	1080	1000	1000	1:36	0.51	0.97	0.67
baseline	30	1080	1000	1000	1:36	0.5	1.0	0.67

Class ratio of 1:1

We use CIFAR-100 to demonstrate the performance of eGAN on a balanced dataset. We notice that the experimental performance follow the same pattern as imbalanced dataset. Training starts with mostly all the majority correctly classified and all the minority misclassified. At epoch 48, a nash equilibrium (with threshold less than or equal to 5) is achieved, with 76 and 71 of minority and majority correctly classified respectively. The maximum F1-score of 0.78 is reached at epoch 53 as shown in Table 3.5. Instead of using 500 samples each of minority and majority class, by training on a single instance of minority and majority sample (*1:1) of CIFAR-100, we obtained an F1-score of 0.63. This demonstrates the impact of transfer learning on the training.

Class ratio of 1:36

For pneumonia subset of CheXpert dataset with imbalance ratio 1:36, the best performed model achieves 0.51, 0.97, and 0.67 precision, recall, and F1-score respectively. The results shown in Table 3.6 is evaluated on 1000 of each minority and majority test set. As can be seen in the table, our approach did not beat the baseline classification model because this task is more of an anomaly detection task rather than a classification problem. Also, the pre-trained image classification model source dataset (i.e. ImageNet) is different from the medical domain. Exploring more variants of complex GAN architectures like BigGAN, StyleGAN and ProGAN could possibly help.

3.5 Conclusion

In this work, we demonstrate the capability of a GAN-based unsupervised technique to address class imbalance using pre-trained models. We conducts experiment with varying levels of imbalance ratios in the training dataset. Instead of synthesizing artificial images with the generator for data augmentation, we employ the discriminator as a classifier and formulate the loss function accordingly. Experimental results reveal that transfer learning plays a significant role in the model performance. The performance measure of interest plays a significant role in deciding the trained model from which epoch to deploy in production as the model at different epochs favour different evaluation metrics such as sensitivity. Future work will focus on the usage of this approach for anomaly detection task where the distinguishing features between normal (majority) and abnormal (minority) are less profound. Our work can be further explored in object detection tasks in case of imbalance between foreground and background.

Chapter 4

CONTRASTIVE LEARNING APPROACH TO ADDRESSING ZERO-SHOT LEARNING PROBLEM

4.1 Introduction

Zero-shot learning (ZSL) is a problem in machine learning, where during test time, a learner observes samples from classes not observed during training when the learner needs to predict sample class[55].

Deep learning (DL) requires large amounts of carefully labeled data, which is very often difficult to acquire and expensive to annotate. Even with large amounts of data, supervised learning still has blind spots in learning useful and rich representations. Moreover, the supervision signal can bias the network in unexpected ways. Self-supervised learning (SSL) is an effective way to extract training signals from massive amounts of unlabeled data and to learn good representation to facilitate downstream tasks where it is expensive to collect task-specific labels. This representation is a feature vector that contains semantic information

about the input image. SSL is a special type of representation learning that enables learning good data representation from unlabeled dataset. SSL is necessary because data labeling is often expensive and thus high-quality labeled datasets are limited. Also, a good learned representation enables easier transfer of useful information/signals to a variety of downstream tasks. In our research, the downstream tasks consists of zero-shot and few-shot transfer to new tasks.

The use of contrastive loss as the cost function is critical to the success of SSL. This objective function pulls together embeddings of augmentations of the same bumblebee species pair and pushes embeddings of views of different species pairs far apart. This is arguably a challenging task, even to humans, because many species of bumblebees are quite similar, even to experts.

Pretext tasks shown in Figure 4.2 are an important component of any SSL training pipeline. Our goal was to have a model learn on downstream tasks like ZSL although models trained on pretext tasks were transferred for specific tasks. On some occasions, pre-training data could be the same as the task-specific data. In our research, we explored two options/approaches. Our pre-training data could be either bumblebees images or a CIFAR-10 dataset. In recent years, researchers have become more creative at determining different forms of pretext tasks. In[56], the task was to have the model predict the degree of rotation in an image. After a patch,[57] predicted which of eight neighboring locations had another patch. This forces the model to learn the innate relationships among patches of the input image. In our research, we formulated our pretext task as a measure of similarity in the same species and dissimilarity between different species of bumblebees.

Existing image-text retrieval models are often fine-tuned on pretext tasks in which the data distribution shift is minimal, but the domain of entomology has terms that are rare in most pre-trained tasks.

The Siamese neural network (SNN) is an architecture that comprises two identical neural networks with shared weights; their parameters are trained to determine the similarity between two data samples using a distance metric on the embeddings. Our research used SNN to determine if two bumblebee images belong to the same specie.

4.2 Related Work

Recently, SSL has gained a lot of momentum as a new paradigm of representation learning because supervised learning often requires a large number of labels, which is time consuming or not available in certain domains. As a result, researchers have become more creative: using part of the data itself as a form of supervision, a seemingly frivolous exercise that has produced remarkable outcomes.

Self-supervision is often achieved either through self-prediction or contrastive learning (CL). In[58], video colorization as a self-supervised learning method provides a rich representation that can be used for video segmentation and unlabeled visual region tracking without extra fine-tuning. Without using supervised labels for training, the zero-shot CLIP[59] classifier performs quite well on challenging image-to-text classification tasks.[60] learned a joint visual-language representation using CL and training on one billion noisy image alt-text pairs. To the best of our knowledge, such an enormous dataset is not yet available in entomology, which makes the learning task harder.

CL has helped make many SSL applications successful. Ideally, CL keeps positive pairs close to each other and negative pairs far apart in the feature space. According to[61;62], CL currently performs at state-of-the-art level in SSL.[63] adopted a memory bank to store the instance class representation vector. Momentum Contrast (MoCo)[64] trained a visual representation encoder by matching an encoded query q to a dictionary of encoded keys with contrastive loss. SimCLR[61] used the normalized temperature-scaled cross-entropy loss as contrastive loss. BYOL[62] relied only on positive pairs, not negative pairs, to learn feature representation. SimSiam[61;64] removed the momentum encoder and provided a simpler design based on BYOL.

In the first paper in which SNN was published,[65], the author depicted the significance of the network for the signature verification task. The goal was to identify signature authenticity.

ZSL is a model's ability to detect classes never seen during training. Even for humans, this is difficult. These classes are sometimes referred to as Out Of Distribution (OOD) samples.

Earlier work in ZSL use attributes in a two-step approach to infer unknown classes. In vision research, more recent, advanced models learn mappings from image feature space to semantic space.[66] proposed a TF-VAEGAN to address generalized zero-shot recognition. In that research, both visual features and corresponding semantics are fed into the architecture. In contrast, we based our approach on only visual signals.

4.3 Proposed Approach

To determine how well our model used a pretext task in a downstream task of novelty detection, we trained our model not on an output class of bumblebees but rather the similarities and dissimilarities between different species using contrastive loss as shown in Figure 4.5. This task forces the model to be aware of visual signals needed in the downstream task. Figure 4.1 shows our convolutional SNN, which uses feature extractors pretrained on ImageNet.

The SNN has three convolutional layers with 16, 32, and 64 filters. We use 3*3 kernel size with the same padding. The final fully connected layer has 128 neurons on which the energy function is computed to estimate the distance vector. Euclidean distance is used as the energy function. Rectified linear activation function (ReLU) is applied after each conv layer followed by maxpooling2D. To avoid overfitting, we applied dropout with probability of 20% after the feature extraction stage. Before passing the image pairs to the feature extractors, we augmented the images by randomly flipping and rotating. We use contrastive loss as the objective function to be optimized. The loss equation is defined in Equation 4.1:

$$L(Y, B1, B2, W) = (1 - Y) \cdot \frac{1}{2} D_w^2 + Y \frac{1}{2} (max(0, m - D_w))^2$$

$$D_w^2 = (f(B1) - f(B2))^2 \tag{4.1}$$

$(B1, B2)$ is a pair of bumblebee species images; D_w^2 is the Euclidean distance which represents the energy function; Y is 1 if the pair is similar and 0 if not; and m represent the margin,

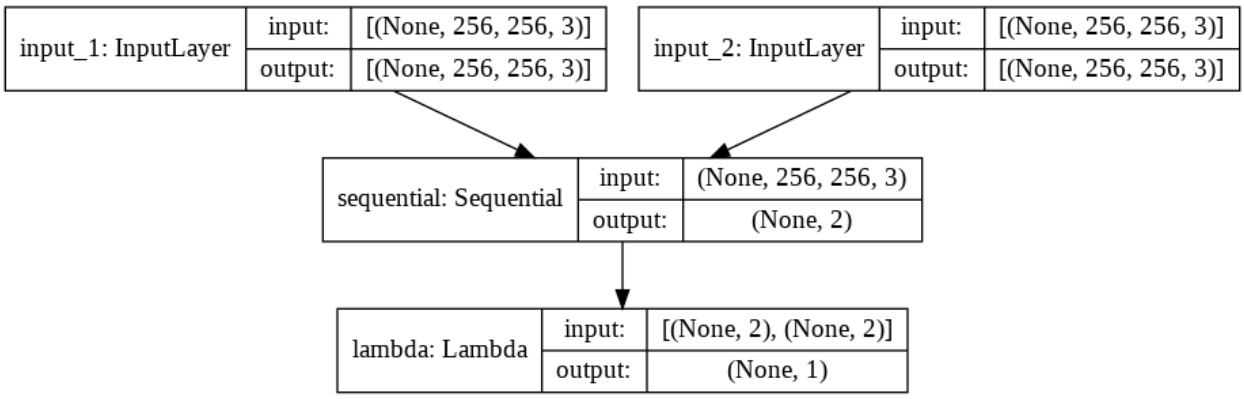


Figure 4.1: *Siamese Neural Network*

a threshold for which the embedding is satisfactorily separable. Finally, W are the network parameters to be learnt during training.

We experimented with MobileNetV2 and ResNet-101 as feature extractors before measuring the energy function between each embedding of the input image pair.

4.4 Methodology and Experimental Design

4.4.1 Dataset

The dataset consisted of 45 species of bumblebee totalling 104770 samples as shown in Table 4.8. Figure 4.3 shows the percentage of the different species in an histogram. Bombus impatiens dominates with 9.6% of the entire dataset.

4.4.2 Data Preprocessing

We preprocessed the dataset by moving all the species with fewer than a thousand samples to the test set, which means that 21 species are not represented in the training set. We then randomly selected 20% of samples of each remaining species to include in the test data. All 45 species are represented in the test data.

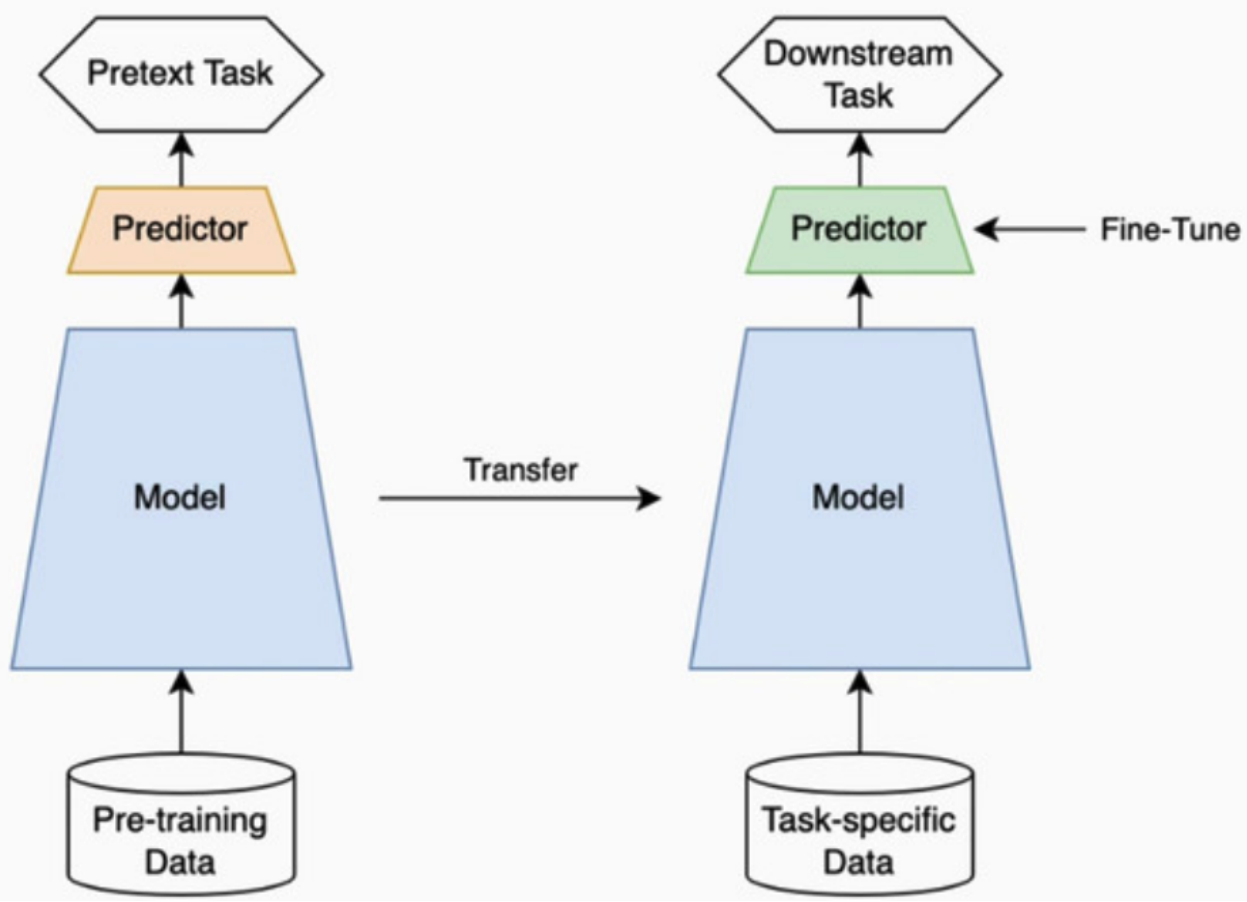


Figure 4.2: *Pretext Tasks*

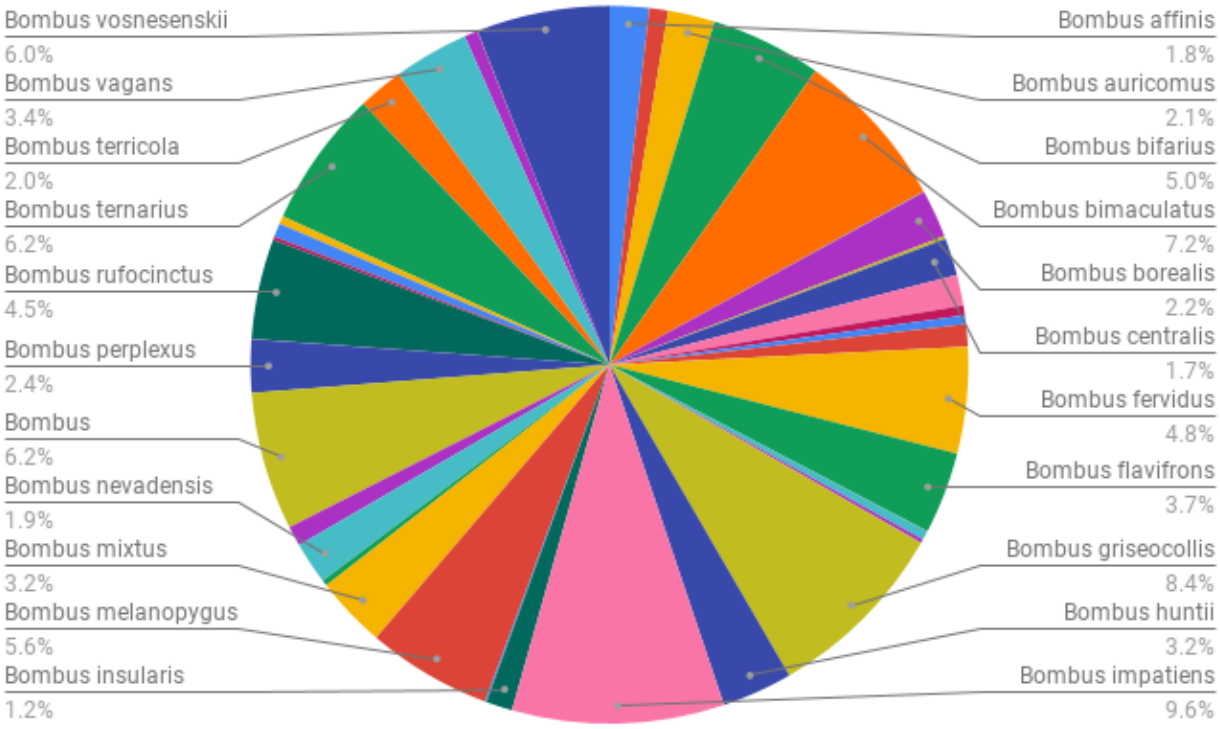


Figure 4.3: *Pie chart of the various species*



Figure 4.4: *Example images of various bee species*

Table 4.1: *Bumblebee (B.) Species.* (***Zero and few shots are in bold***)

Species	#samples	Species	#samples
B. appositus	**888**	B. affinis	1855
B. crotchii	**457**	B. bifarius	5237
B. bohemicus	**21**	B. bimaculatus	7595
B. caliginosus	**147**	B. borealis	2256
B. franklini	**9**	B. citrinus	1449
B. cockerelli	**3**	B. centralis	1755
B. cryptarum	**419**	B. fernaldae flavidus	1047
B. suckleyi	**14**	B. flavifrons	3866
B. fraternus	**424**	B. fervidus californicus	5020
B. frigidus	**179**	B. griseocollis	8787
B. variabilis	**12**	B. impatiens	10008
B. jonellus	**5**	B. insularis	1241
B. kirbiellus	**69**	B. melanopygus	5892
B. morrisoni	**276**	B. mixtus	3359
B. natvigi	**3**	B. nevadensis	1982
B. occidentalis	**943**	B. pensylvanicus sonorus	6522
B. polaris	**11**	B. perplexus	2503
B. sandersoni	**163**	B. rufocinctus	4719
B. sitkensis	**647**	B. huntii	3335
B. sylvicola	**366**	B. ternarius	6467
B. vandykei	**625**	B. terricola	2145
		B. vagans	3546
		B. vosnesenskii	6284
		B. auricomus	2219
		Total = 104770	

Table 4.2: *ResNet-101 confusion matrix at 0.5 threshold on testset-ALL (45 species)*

	Predicted similar	Predicted dissimilar
Actually similar	61	29
Actually dissimilar	43	47

4.4.3 Experimental results

We trained a convolutional SNN for a maximum of 100 epochs, with patience level of 7, on 80% of the training set. We then evaluated the trained model on the 20% validation set. The results are shown in Table 4.9.

Table 4.3: *ResNet-101 confusion matrix at 0.5 threshold on testset (24 species)*

	Predicted similar	Predicted dissimilar
Actually similar	3725	1267
Actually dissimilar	1514	3478

Table 4.4: *ResNet-101 Confusion matrix at 0.5 threshold on 21 unknown test set species - zero-shot learning*

	Predicted similar	Predicted dissimilar
Actually similar	24	18
Actually dissimilar	15	27

4.5 Results and Discussion

Table 4.9 shows the average accuracy, recall, precision, and F1 score at different thresholds/cut-offs. Of all the thresholds, the results reveal that 0.5 is a good decision boundary for most species. Similarity scores below the threshold are classified as similar and vice versa. We also inspected the performance of contrastive learning within each species of the bumblebees to discover hard and easy data pairs. On the test data, the best performing SNN model, which uses ResNet-101 as base feature extractor, achieves 73% precision, 72% recall, 73% F1-score, and 72% accuracy.

The results shown in Table 4.10 show that our SSL network can detect zero shot instances of the bumblebees although the SNN performs much better on species that share intersection with the training dataset. Table 4.5 shows that our network can still differentiate between pairs of zero-shot instances. The network could differentiate the *B. cockerelli* and *B. fraternus* pair as well as the *B. cryptarum* and *B. fraternus* pair with a high confidence. This is useful in deciding in what species the novel instance falls. With this, an entomologist can quickly

Table 4.5: *F1-scores of zero-shot species pair*

	B. natvigi	B. variabilis	B. appositus	B. fraternus	B. frigidus
B. cockerelli	0.5	**0.88**	0.28	**1.00**	**1.00**
B. polaris	0.5	0.43	0.50	0.72	0.46
B. occidentalis	0.11	0.48	0.52	0.60	0.63
B. cryptarum	0.62	0.49	0.51	**0.74**	0.58
B. caliginosus	0.50	0.39	0.52	0.60	0.59

Table 4.6: *ResNet-101 confusion matrix at 0.5 threshold on validation*

	Predicted similar	Predicted dissimilar
Actually similar	3197	1123
Actually dissimilar	1243	3077

Table 4.7: *MobileNetV2 confusion matrix at 0.5 threshold on validation*

	Predicted similar	Predicted dissimilar
Actually similar	2897	1423
Actually dissimilar	1608	2712

narrow down a novel species class to either in-distribution or OOD samples.

4.5.1 Conventional Zero-shot learning

This involves learning a classifier that can detect only unknown test samples. Our network achieved 61% F1-score on 5681 unknown species of bumblebees. This task is made more difficult because of a high class imbalance with some species having very few samples, as can be seen in Table 4.1.

4.5.2 Generalized Zero-shot learning

The goal of generalized zero-shot learning is to train a classifier that can detect test instances from seen or unseen classes. This is more useful, and thus more challenging. With the entire 45 species in the test set, our network achieved a 60% F1-score on various pairing of the species as shown in Table 4.10.

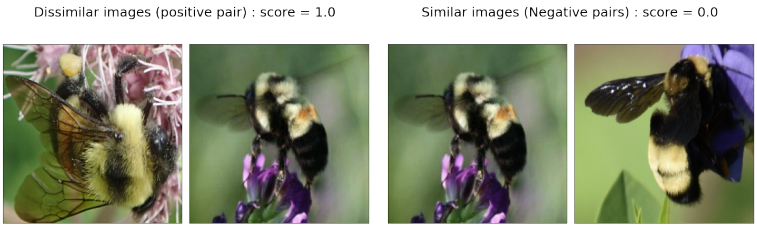


Figure 4.5: *Positive pair and Negative pair*

Table 4.8: *Number of samples in the dataset.*

Dataset	images	species-represented
Train + Validation	79280	24
Test	25490	45

4.6 Conclusion

In this work, we explored using a contrastive learning approach to detect novel bumblebee instances in a zero-shot setting. We showed that our model performs well in learning the embeddings of similar and dissimilar bumblebee species pairs. Our experimental results showed that our model can generalize to zero shot instances that were not part of the training. Future work will focus on improving the performance on the ZSL task by learning a multi-modal network that maps images to text-attribute descriptions. We can also explore cosine similarity as the energy and triplet loss as the loss function.

Table 4.9: *Classification Accuracy, F1, Precision, Recall when using different methods on validation set.*

Models	Threshold	Precision	Recall	F1	Accuracy
SNN + ResNet-101[22]	0.5	0.73	0.72	0.73	0.72
SNN + MobileNetV2[29]	0.1	0.66	0.51	0.35	0.50
	0.2	0.62	0.55	0.47	0.55
	0.3	0.63	0.60	0.58	0.60
	0.4	0.63	0.63	0.64	0.63
	0.5	0.65	0.65	0.65	0.65
	0.6	0.65	0.65	0.64	0.65
	0.7	0.67	0.66	0.65	0.66
	0.8	0.69	0.65	0.63	0.65
	0.9	0.69	0.62	0.58	0.62

Table 4.10: *Classification Accuracy, F1, Precision, Recall on test set using ResNet-101 as feature extractor.*

	Threshold	Precision	Recall	F1	Accuracy
21 species (Zero-Shot)	0.5	0.61	0.61	0.61	0.61
45 species (ALL)	0.5	0.61	0.60	0.60	0.60
24 species	0.5	0.73	0.72	0.72	0.72

Chapter 5

CONCLUSION

5.1 Summary and Review of Contributions

This research work is focused on developing deep learning techniques that can automatically detect anomalies in image data by using computer vision. This task involves: learning the subtle difference between normal and abnormal class, class imbalance and Zero-shot learning. Methods developed include: AdeNet, eGAN and Siamese Neural Network.

Firstly, we address the task of detecting damaged electrical insulators in power line images, an important safety task. Detecting damaged electrical insulators has long been a labor-intensive task requiring substantial experience and careful examination of the images.

The new architecture, AdeNet, based on a deep convolutional neural network has the advantage of requiring little energy, allowing it to be used on low-energy devices like UAVs, which are often used to acquire the data for identifying damaged insulators. The low energy allows AdeNet to make decision on the UAV, which can then be used to acquire more data of suspect insulators during a mission, without going back to base and re-deploying the UAV. Like intuition in[37;38], we found an architecture that is not as simple as LeNet-5 and not as complex as ResNet (just the right mix of simplicity and complexity) that also performs well in detecting anomalies in damaged insulators in power lines.

In comparing the proposed method and other solutions, including shallow learning, we

found that deep learning architectures outperform shallow learning architectures. Within DL architecture, fine-tuning the pre-trained CNN models did not improve the performance in all cases. We demonstrated this through experiments with our AdeNet architecture. Experimental results also showed that our architecture, without pre-training, outperformed all other deep neural networks. This makes AdeNet suitable for mobile/embedded devices that have storage and computation constraints.

Secondly, we demonstrate the capability of a GAN-based unsupervised technique to address class imbalance using pre-trained models. We conduct experiment with varying levels of imbalance ratios in the training dataset. Instead of synthesizing artificial images with the generator for data augmentation, we employ the discriminator as a classifier and formulate the loss function accordingly. Experimental results reveal that transfer learning plays a significant role in the model performance.

The performance measure of interest plays a significant role in deciding the trained model from which epoch to deploy in production as the model at different epochs favour different evaluation metrics for example sensitivity.

Lastly, we explore using contrastive learning approach to detect novel bumblebee instances in a zero-shot setting. We showed that our model performs well in learning the embeddings of similar and dissimilar bumblebee species pairs. Our experimental results showed that our model can generalize to zero shot instances that were not part of the training. Future work will focus on improving the performance on the ZSL task by learning a multi-modal network that maps images to text-attribute descriptions. We can also explore cosine similarity as the energy and triplet loss as the loss function.

5.2 Future Directions

It is important to remember that when machine learning systems are deployed to production use, it is not merely optimizing a predictive model. Rather, we are providing a tool that will be used to automate, sometimes, critical decisions. The leap from considering predictions to decisions raises not only new technical questions, but also a slew of ethical questions that

must be carefully considered. Despite the remarkable performance of DL-based models in anomaly detection task, there are still a number of failure cases that lacks explainability. As seen in Table 4.5 on zero-shot learning, it will be a good idea to be able to mathematically formulate theories behind the non-performance of the network. Moreover, specialized application areas like medicine, disease diagnosis and safety-critical systems require adequate justification of model's performance in extreme cases before they can be deployed. Of recent, neuro-symbolic AI, a field that seeks to incorporate knowledge-based system and logic into architecting neural network design, seek to provide trust to the neural network's *black boxes*. These will be an interesting area of research focus as a follow up on anomaly detection.

Bibliography

[1] Laura Beggel, Michael Pfeiffer, and Bernd Bischl. Robust anomaly detection in images using adversarial autoencoders. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 206–222. Springer, 2019.

[2] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems*, pages 9758–9769, 2018.

[3] Desire Sidibe, Shrinivasan Sankar, Guillaume Lemaitre, Mojdeh Rastgoo, Joan Massich, Carol Y Cheung, Gavin SW Tan, Dan Milea, Ecosse Lamoureux, Tien Y Wong, et al. An anomaly detection approach for the identification of dme patients using spectral domain optical coherence tomography images. *Computer Methods and Programs in Biomedicine*, 139:109–117, 2017.

[4] Qi Wei, Yinhao Ren, Rui Hou, Bibo Shi, Joseph Y Lo, and Lawrence Carin. Anomaly detection for medical images based on a one-class classification. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, page 105751M. International Society for Optics and Photonics, 2018.

[5] Kamal Malik, Harsh Sadawarti, and Kalra G S. Comparative analysis of outlier detection techniques. In *IJCA*, volume 97, pages 12–21, 2014.

[6] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9592–9600, 2019.

[7] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but differnet: Semi-

supervised defect detection with normalizing flows. *arXiv preprint arXiv:2008.12577*, 2020.

[8] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pages 146–157. Springer, 2017.

[9] Thomas Schlegl, Philipp Seebock, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44, 2019.

[10] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018.

[11] Paolo Napoletano, Flavio Piccoli, and Raimondo Schettini. Anomaly detection in nanofibrous materials by cnn-based self-similarity. *Sensors*, 18(1):209, 2018.

[12] Tobias Böttger and Markus Ulrich. Real-time texture error detection on textured surfaces with compressed sensing. *Pattern Recognition and Image Analysis*, 26(1):88–94, 2016.

[13] Van Nhan Nguyen, Robert Jenssen, and Davide Roverso. Automatic autonomous vision-based power line inspection: A review of current status and the potential role of deep learning. *International Journal of Electrical Power & Energy Systems*, 99:107–120, 2018.

[14] François Mirallès, Nicolas Pouliot, and Serge Montambault. State-of-the-art review of computer vision for the management of power transmission lines. In *Proceedings of the 2014 3rd International Conference on Applied Robotics for the Power Industry*, pages 1–6. IEEE, 2014.

[15] Xian Wang and Youmin Zhang. Insulator identification from aerial images using support

vector machine with background suppression. In *International Conference on Unmanned Aircraft Systems*, pages 892–897. IEEE, 2016.

[16] Shenglong Liao and Jubai An. A robust insulator detection algorithm based on local features and spatial orders for aerial images. *IEEE Geoscience and Remote Sensing Letters*, 12(5):963–967, 2014.

[17] Jingjing Zhao, Xingtong Liu, Jixiang Sun, and Lin Lei. Detecting insulators in the image of overhead transmission lines. In *International conference on intelligent computing*, pages 442–450. Springer, 2012.

[18] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[19] Xiren Miao, Xinyu Liu, Jing Chen, Shengbin Zhuang, Jianwei Fan, and Hao Jiang. Insulator detection in aerial images for transmission line inspection using single shot multibox detector. *IEEE Access*, 7:9945–9956, 2019.

[20] R Ishino and F Tsutsumi. Detection system of damaged cables using video obtained from an aerial inspection of transmission lines. In *IEEE Power Engineering Society General Meeting, 2004.*, pages 1857–1862. IEEE, 2004.

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[23] Yue Liu, Jun Yong, Liang Liu, Jinlong Zhao, and Zongyu Li. The method of insulator recognition based on deep learning. In *2016 4th International Conference on Applied Robotics for the Power Industry (CARPI)*, pages 1–5. IEEE, 2016.

[24] Lior Shamir. Udat: A multi-purpose data analysis tool. *ASCL*, pages ASCL–1704, 2017.

[25] Lior Shamir, Nikita Orlov, D Mark Eckley, Tomasz Macura, Josiah Johnston, and Ilya G Goldberg. Wndchrm–an open source utility for biological image analysis. *Source code for biology and medicine*, 3(1):13, 2008.

[26] Nikita Orlov, Lior Shamir, Tomasz Macura, Josiah Johnston, D Mark Eckley, and Ilya G Goldberg. Wnd-charm: Multi-purpose image classification using compound image transforms. *Pattern Recognition Letters*, 29(11):1684–1693, 2008.

[27] Lior Shamir. Evaluation of face datasets as tools for assessing the performance of face recognition methods. *International Journal of Computer Vision*, 79(3):225, 2008.

[28] Lior Shamir, Tomasz Macura, Nikita Orlov, D Mark Eckley, and Ilya G Goldberg. Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. *ACM Transactions on Applied Perception*, 7(2):1–17, 2010.

[29] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[30] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[31] Yann LeCun et al. Lenet-5, convolutional neural networks. *URL: http://yann. lecun. com/exdb/lenet*, 20(5):14, 2015.

[32] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[35] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*, 2020.

[36] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[37] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. *arXiv preprint arXiv:1904.03626*, 2019.

[38] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[39] Chaoyun Zhang, Pan Zhou, Chenghua Li, and Lijun Liu. A convolutional neural network for leaves recognition using data augmentation. In *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, pages 2143–2150. IEEE, 2015.

[40] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[41] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[42] Yanmin Sun, Mohamed S Kamel, Andrew KC Wong, and Yang Wang. Cost-sensitive

boosting for classification of imbalanced data. *Pattern recognition*, 40(12):3358–3378, 2007.

[43] Charles X Ling and Victor S Sheng. Cost-sensitive learning and the class imbalance problem. *Encyclopedia of machine learning*, 2011:231–235, 2008.

[44] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *Proceedings of the European Conference on Computer Vision*, pages 218–234, 2018.

[45] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pages 1–8. Citeseer, 2003.

[46] Nasik Muhammad Nafi and William H Hsu. Addressing class imbalance in image-based plant disease detection: Deep generative vs. sampling-based approaches. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 243–248. IEEE, 2020.

[47] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[48] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[49] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

[50] Yangru Huang, Yi Jin, Yidong Li, and Zhiping Lin. Towards imbalanced image classification: A generative adversarial network ensemble learning method. *IEEE Access*, 8: 88399–88409, 2020.

[51] Alessandro Bria, Claudio Marrocco, and Francesco Tortorella. Addressing class imbalance in deep learning for small lesion detection on medical images. *Computers in Biology and Medicine*, 120:103735, 2020.

[52] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[53] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[54] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

[55] Wikipedia contributors. Zero-shot learning — Wikipedia, the free encyclopedia, 2021. URL https://en.wikipedia.org/w/index.php?title=Zero-shot_learning&oldid=1018108538. [Online; accessed 26-April-2021].

[56] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[57] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[58] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 391–408, 2018.

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al.

Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[60] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

[61] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[62] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[63] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[64] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[65] Jane Bromley, James W Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688, 1993.

[66] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In

Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16, pages 479–495. Springer, 2020.