CS230

Active Learning to Solve Class Imbalance in Bird Species Classification

Christian Gabor
Department of Computer Science
Stanford University
gaborc@stanford.edu

Abstract

Bird species serve as an important indicator for habitat health in ecology. With the advance in computer vision through deep learning, this paper explores the application of convolution neural networks to bird species classification. This paper also explores the effect of active learning on unbalanced datasets to improve model performance through selective sampling of which images to label for training. Both maximum entropy and least confidence active learning improved the model performance to create a more robust network.

1 Introduction

Species classification is an important domain in ecology and wildlife preservation as bird serve as a bio indicator [3]. Currently, many species are going extinct each year, and having accurate methods for documenting these changes can help ecologists come up with solutions. Birds are not only endangered, but can also serve as a bio indicator for ecological health. However, because endangered species are more likely to be mislabeled by non-experts simply due to real world frequency imbalance, it is important to solve this issue of class imbalance to deploy robust applications. For this project, I will be exploring how to mitigate class imbalance in bird image datasets by using an active learning approach to prioritize which birds among unlabeled datasets should be labeled.

Industrial deep learning applications for many target domains also run into the issue of imbalanced datasets. Because of this imbalance, important classes may be misclassified. For instance, in bird species recognition the model may misclassify rare species as a more common visually similar species. Annotating this rare species could be challenging as a professional ornithologist may need to label the correct species rather than the common birdwatchers who may make mistakes. In many domains, the class imbalance leads to inferior model performance when the rare class is important to classify correctly. For example, in medical diagnostics many patients are healthy but on the rare case of a disease, it is very important to have good recall because it could be deadly to miss this diagnosis. It is even more expensive in this domain to annotate data because medical experts need to view the data to fix the class imbalance.

In this project, a neural network with a softmax output of 200 classes trains on feature extractions from a pretrained convolutional neural network in transfer learning. The training data of bird species is intentionally down-sampled to simulate an unbalanced dataset. The active learning metric then determines which input images to assign with a bird species label before adding to the training dataset for further training.

CS230: Deep Learning, Fall 2019, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Related work

2.1 Bird Species Classification

Due to the importance of environmental monitoring to study today's changing ecosystems, Chen et al. explore deep convolutional neural networks to predict common North American animal species [2]. They show that CNNs outperform other methods on species classification when recognizing animals from camera trap systems located in remote wilderness settings. They train their model with 14346 images for 20 output class labels. Because this research classifies common species, it is not likely their dataset suffered from class imbalance, although this is not emphasized in the article.

2.2 Class Imbalance

In the deep learning community, the most common technique to solve class imbalance issues is to over sample lower frequency classes, or to add a weighted learning rate function to account for this in backpropagation update step [1]. There has been little research in recent years as deep learning has taken off as to the performance penalty of class imbalance in deep learning regimes. Buda et al. find that class imbalance detrimentally impacts the performance of the ROC curve or precision and recall of deep learning models [1]. This issue of class imbalance has been studied extensively for other machine learning paradigms, but the current lack of literature suggests it is therefore important to study class imbalance for newer deep learning models that will be used in domains with class imbalance.

2.3 MobileNet V2

MobileNet V2 is an improvement upon MobileNet V1 designed as a low latency convolutional neural network suitable for mobile computing [4]. This network is composed of MobileNet building blocks that perform an efficient depth wise separated convolution with intermediate skip connections. The down sampling representation of these convolutional filters can be used for image recognition tasks such as classification, object detection and segmentation. A pretrained MobileNet V2 model was used for this project that was trained on the ImageNet dataset which contains several existing bird classes such as goose and robin.

2.4 Active Learning

Burr Settles explores various active learning techniques applied to the machine learning field and describes different techniques for implementing active learning [5]. Different techniques involve either pool based learning methods which look at the entire pool of unlabeled data to select optimal points, or to look at individual pool streams and select to label a point if it passes a value threshold. Two techniques were used from this literature survery including maximum entropy and least confidence scores. Many of the machine learning paradigms discussed in this book were applied to traditional machine learning before the newest use of deep learning for image classification tasks and this project explores the application of the methods described in this book to newer deep learning paradigm.

Wang et al. explore active learning methods for deep convolutional neural networks and find that active learning can improve model performance [7]. This study, however, does not focus on unbalanced datasets and instead looks at increasing efficiency of active learning techniques for datasets that have high confidence scores on existing classes.

3 Dataset and Features

This project evaluates the performance of transfer learning of MobileNet V2 on the Caltech-UCSD Bird-200-2011 dataset[6], composed of 11,788 bird images over 200 species.

The dataset is broken into 80 percent training data and 20 percent test data. To test class imbalance with active learning the training dataset is further split into common classes and rare classes, where 100 classes are down-sampled to only have 10 percent training images. Thus, only half of the bird species are fully represented in the training data.

As shown in figure 1, MobileNet V2 is used to extract features from these images which reduced the input size of the network from 224x224x3 to 1080 features. MobileNet V2 is pretrained on ImageNet-1000 which contains general bird class labels, such as goose and robin. It is therefore likely that the feature maps were pretrained to pick out bird features that could help in transfer learning to a new 200-class species classification task. To speed up training, the MobileNet V2 network was frozen and training was done on the new output layers.

The active learning component uses the 200 feature softmax output of the trainable neural network to select data to add to the training dataset and is described in more detail in further sections

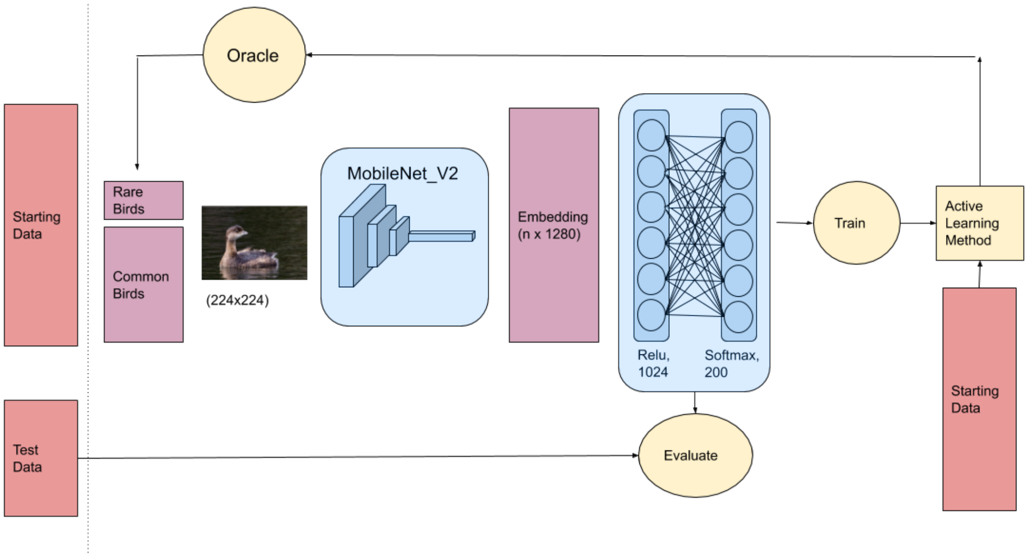

Figure 1: Data Pipeline

Figure 2: Caltech-UCSD Bird-200-2011 Dataset

4 Methods

Before active learning was applied, a baseline model was trained on the MobileNet V2 features and hyper-parameters were selected on a validation dataset held out from the training dataset. The features pass through a trainable neural network with one hidden layer of 1024 neurons before being applied to a 200 neuron soft max classification output layer. Dropout of 0.5 was applied before both layers to avoid over fitting. L1 and L2 normalization was tested, but these hyper-parameters did not improve validation performance. This model was trained for 30 epochs on the training dataset before active learning was applied.

With the classification pipeline in place, active learning was tested on the full training dataset before another epoch of training. Two active learning techniques were tested: maximum entropy and least confidence.

$$MaximumEntropy(X) = max_{x^{[i]} \in X} \left(-\sum_{j} p_\theta(y_j|x^{[i]}) log_2(p_\theta(y_j|x^{[i]})) \right)$$

Maximum Entropy effectively computes the Shannon Entropy Rate of each softmax output of the classification network and the active learning metric aims to find the maximum across the entire dataset of unlabeled samples. The inner term of this equation multiplies the prediction class probability with the log of that probability to compute the entropy rate, and the active learner looks for the max over X samples. The idea behind this algorithm is that when the model has high uncertainty in the class label, which will be likely for the rare classes, the active learning component will select classes to reduce this uncertainty.

$$LeastConfidence(X) = min_{x^{[i]} \in X} \left(max_{y_j^{[i]}} (p_\theta(y_j|x^{[i]}) \right)$$

Least confidence active learning searches for input examples that have the lost top prediction certainty from the softmax prediction. The inner term computes the maximum over the output softmax vector to identify the predicted class label, then the active learning component aims to find the minimum over all unlabeled examples, representing least confidence. The concept behind this approach is to use the uncertainty of the network as an indicator for high variance in the input data.

5 Experiments/Results/Discussion

Method	Accuracy	Precision	Recall
Upper Bound	0.875	0.887	0.875
Random	0.729	0.788	0.729
Max Entropy	0.826	0.857	0.827
Least Confidence	0.830	0.847	0.830

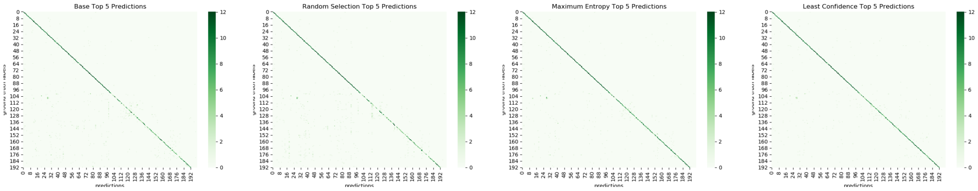

Figure 3: Confusion Matrices in order: 1. Before Active Learning 2. Random Selection 3. Max Entropy 4. Least Confidence

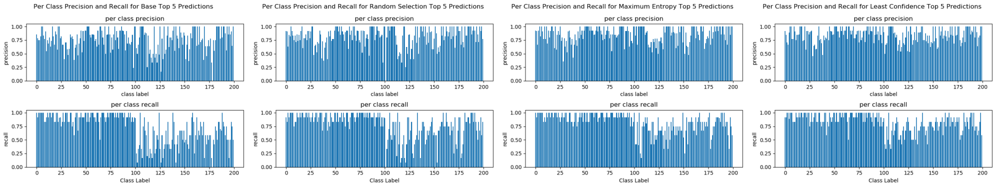

Figure 4: Precision and Recall for each class in order: 1. Before Active Learning 2. Random Selection 3. Max Entropy 4. Least Confidence

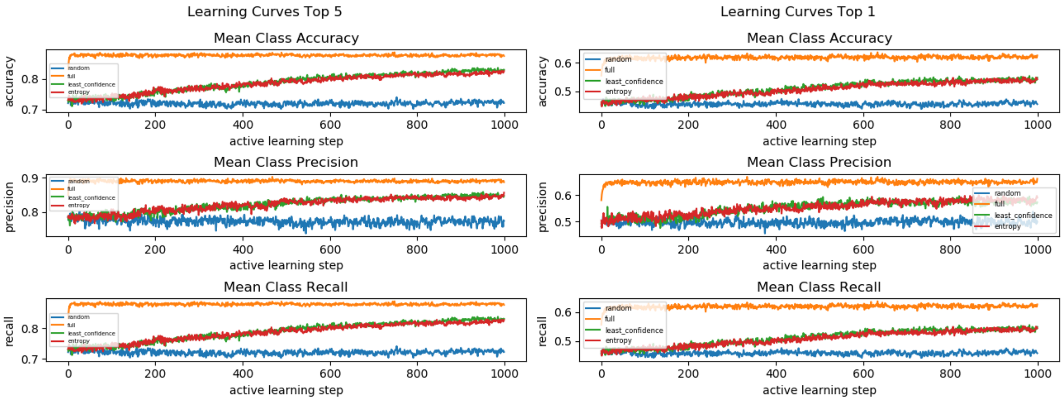

Figure 5: Effects of Active Learning on Test Performance. In each graph, the top curve represents upper bound performance with all labels. The other three curves represent max entropy, least confidence and random selection.

A baseline was performed with 30 epochs to establish performance of the model before active learning. It is clear to see that the 100 under sampled classes have poor recall performance. In the confusion matrices, the lower classes are less likely to be found in the test dataset, and this negatively impacts the precision of the common classes. It is important to note that during training, oversampling the rare classes was still present in the update gradient step which is the common approach to solving class imbalance in deep learning models.

After 1000 epochs of active learning selecting 1 example from the pool of unlabeled samples, both maximum entropy and least confidence scored equally. To compare this to a baseline, random selection was performed on the pool of training data to compare how a naive labeling approach would impact model performance. To get an upper bound on model performance, the model was also trained on the entire dataset with all labels available. This situation would not be likely in a real world situation, but gives the maximum possible performance of the model with the available data. Recall improved the most, which computes $\frac{TruePositives}{FalseNegatives}$. This is likely because the prior network was assigning the true rare classes falsely to the common classes due to imbalance.

6 Conclusion/Future Work

Active learning with least confidence increased the baseline accuracy from 0.72 to 0.83 and maximum entropy increased from 0.72 to 0.826. It is interesting to see how active learning performed much better than a naive random selection approach. This could imply that research teams that do not optimize their data labeling strategy could suffer from inferior model performance compared to an active learning approach.

This specific project could be applied to a larger effort in environmental conservation through citizen science. For those interested in conservation but may know little about bird species may take pictures of birds when they go on walks at local parks. They could upload these unlabeled bird images to an archive that extracts the image metadata about time and location. This archive may contain millions of bird images, but there are only so many expert ornithologists that can accurately label the bird images that an active learning approach must be used to select which images to show to experts. As the ornithologists label data, this data can help improve model accuracy for remote sensors and distributed AI that monitors the environment for ecological conservation efforts.

The next steps on the project would be to apply this active learning task to a larger dataset with more variance in the data. Many of the images in this dataset had birds centered in the image frame, but to assess performance in real data it would be important to experiment with higher variance data. Further research could also look at the impact of unfreezing the pretrained network if computational resources were available. In addition, more sophisticated active learning techniques could be explored to assess model performance, as the literature on active learning for deep neural networks remains sparse.

7 Contributions

The MobileNet v2 feature extractor was taken from Tenforflow hub. This model can be found on github at this link: `https://github.com/tensorflow/models/tree/master/research/slim/nets/mobilenet`

The dataset for Caltech-UCSD Birds-200-2011 can be found at `http://www.vision.caltech.edu/visipedia/CUB-200-2011.html`

This project was done for Stanford CS 229 with no additional team members.

8 Code

The source code for this project can be found at the following link: `https://github.com/gaborchris/ActiveLearningBirds`

This project was implemented in Python 3.6 using Tensorflow 2.

References

[1] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249 – 259, 2018.

[2] Guobin Chen, Tony Han, Zhihai He, Roland Kays, and Tavis Forrester. Deep convolutional neural network based species recognition for wild animal monitoring. *2014 IEEE International Conference on Image Processing, ICIP 2014*, pages 858–862, 01 2015.

[3] Timo Pakkala, Andreas Lindén, Juha Tiainen, Erkki Tomppo, and Jari Kouki. Indicators of Forest Biodiversity: Which Bird Species Predict High Breeding Bird Assemblage Diversity in Boreal Forests at Multiple Spatial Scales? *Annales Zoologici Fennici*, 51(5):457 – 476, 2014.

[4] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[6] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011.

[7] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016.