Active Learning for ConvNets in Bird Species Classification

Name: Christian Gabor
SUNet ID: gaborc

Introduction

How often have you seen a bird on a nice walk and wondered what species that bird was? This neural network built off a pre trained MobileNet_V2 is here to answer your question!. However, there is one problem: bird species distributions are not uniform and the training data is greatly imbalanced. You don't want the network to give you the most common species when you encounter a rare and exciting species.

Many real domains suffer from imbalanced data labels in ground truth data. Labeling data to correct the imbalance is time consuming and expensive. In domains such as healthcare this could require knowledge from a highly paid doctor. In this domain of bird classification we need expert ornithologists to label the data.

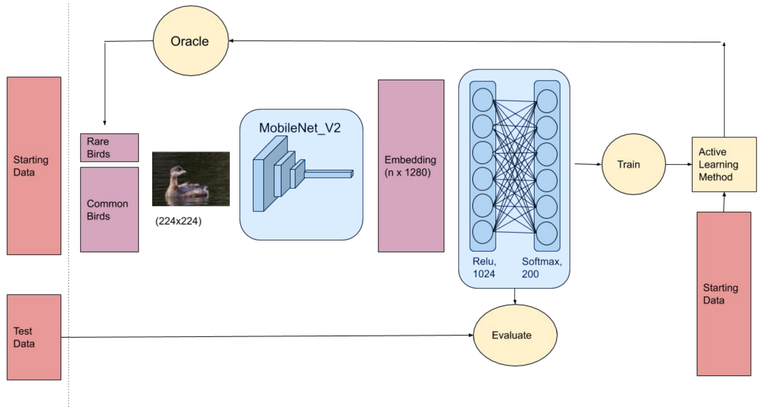
To combat this problem, this project explores active learning which selects the optimal data point to label and add to the dataset.

Data

Dataset
Data for this project taken from the Caltech-UCSD Birds-200-2011 set, which includes 200 species and 11,788 images with roughly equal distribution of 60 images for each species. Before training, a test set of 20% of the original data was set aside.

Simulated Distribution
To simulate the imbalanced dataset, the training data was split into two halves, where the second half of the data was sampled by a factor of 0.1. The remaining data was left for the active learning component.

Pipeline



Training Procedure:
1. Setup datasets for experiments
2. Pretrained MobileNet_V2 predicts embeddings for training set
3. New model head predicts softmax class probabilities
4. Active learning selects highest value x value to query
5. Oracle assigns class label and appends to dataset

For step 5, a human would fill the role of the oracle to provide the label; however, for the purposes of this experiment the label is provided by the ground truth labels.

Methods

$$LC(X) = min_{x_i \in X}(max(softmax(x_i)))$$

Least confidence takes each softmax prediction on the unlabeled x input and assigns the optimal choice to the one with the lowest top_1 prediction.

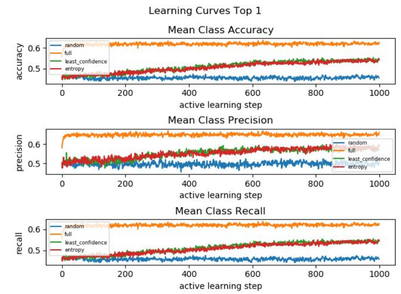
$$max_H(X) = max_{x_i \in X}(-\sum p_\theta(y_j|x_i) log_2(p_\theta(y_j|x_i)))$$

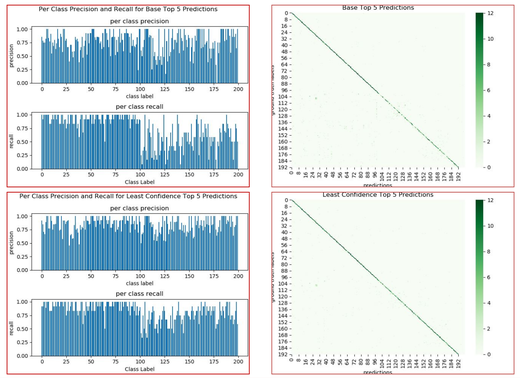
Entropy computes the Shannon entropy rate of each softmax prediction and chooses the example with the highest entropy.

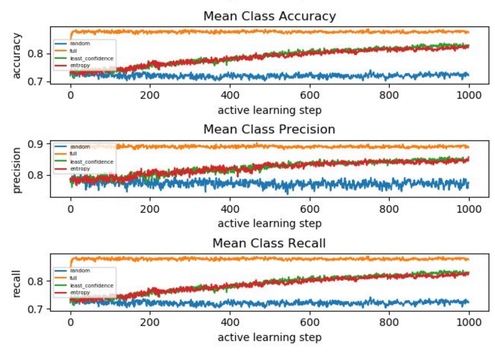
Results

Several observations arise from these experiments
1. The model accuracy improves by over 10% from the baseline accuracy
2. Randomly sampling from the same class distribution did not improve accuracy

This could imply that labeling data at random as it comes in from a data source may slow down model improvement, while active learning could help improve performance with lower labeling effort.







Test Samples