An Exploration of Computer Vision Techniques for Bird Species Classification

Anne L. Alter, Karen M. Wang

December 15, 2017

Abstract

Bird classification, a fine-grained categorization task, is a complex task but crucial in improving and identifying the best computer vision algorithms to use in the broader image recognition field. Difficulties like lighting conditions, complex foliage settings, and similarities in subspecies of birds are just some of the challenges faced by researchers. We implemented softmax regression on manually observed binary attributes, a multi-class SVM on HOG and RGB features from photos, and finally a CNN using transfer learning to classify birds. The pre-trained CNN with fixed feature extraction proved to be the best method for classification with computer vision.

1 Introduction

Recent machine learning and deep learning research has put a lot of focus on facial recognition, which involves detecting faces within images, landmarking specific features to account for the angle at which the picture was taken, and finally determining whose face is in the picture. This is extremely valuable research and has many applications including social media, computer security, marketing, and healthcare. While this research is booming, there has been less focus on the classification of other organisms like birds, for example. As with facial recognition, there is much complexity in the classification of birds due issues like the similarities in different subspecies of birds, challenges with the foreground or background setting of images, and lighting conditions in photos. We intend to learn and classify different bird species from a classified data set that was compiled by Caltech and UCSB. In the process of classifying the birds, we plan to identify specific bird features of interest (eyes, tip of beak, wing, etc).

2 Related Work

Computer vision research on fine-grained categorization of bird species has evolved from human-assisted learning algorithms to algorithms that utilize histogram oriented gradient features (HOG) and scale-invariant feature transforms (SIFT) to more advanced Convolutional Neural Networks (CNNs). Early human-assisted algorithms were clever in that they prompted users to do tasks that were typically difficult for computer vision algorithms, like asking users to do part localization (i.e. identify head, body, etc.), and then allowing learning algorithms to take over, resulting in a decreased need for human input over time [1]. Slightly more advanced algorithms, like Part-Based One-vs-One Feature Recognition (POOF), a method which learned feature alignments but still incorporated human part localization, boosted accuracy to around 50-60% [2]. Some of the first pure computer vision algorithms, without any human input, were able to achieve an accuracy of 10% by using RGB histograms and SIFT inputs [3]. Convolutional neural networks were first implemented with fully supervised learning and their features were iteratively evaluated for their efficacy on computer vision specific features; these algorithms further boosted accuracies to around 65% [4]. The current state-of-the-art CNNs utilize pre-trained networks and are able to normalize bird poses (accounting for one of the biggest challenges of different photo angles and bird positions). These methods have improved accuracies to about 85% [5]. Overall, many computer vision methods still require some human input parameters, but CNNs show the most promise to identifying bird features with pure computer vision and thus improving bird classification accuracy.

3 Dataset

For the project, we will be using the Caltech-UCSD-Birds-200-2011 dataset [3]. This dataset includes 200 categories of bird species, 11,788 total number of images, and other information such as labeled visible bird parts, binary attributes, and bounding boxes surrounding the birds. The authors also include a recommended test and training set split of the data.

4 Methods and Results

We will implement 3 methods and compare them. The first implementation is a softmax Regression of the Binary Attributes, which does not utilize computer vision, but is used as a baseline for bird classification. The second implementation is a multi-class SVM and utilizes computer vision methods to extract HOG gradients and RGB histogram values to classify the birds. And finally transfer learning techniques are used where a pre-trained CNN is loaded with the last three layers modified.

4.1 Softmax Regression on Binary Attributes

The features for this method are 312 bird attributes that have been manually classified by researchers at Caltech and UCSD. These features denote attributes like bill shape, wing color, tail pattern, and overall bird size. All attributes are binary and thus simply represented in a 312x1 feature vector for each training example. Attributes like wing color are broken down into specific wing colors (blue wing, black wing, red wing, etc...) so that there is a simple binary representation.

A regularized softmax algorithm is implemented twice, once on the training data with specifically classified species (ex. Prarie and Pine Warblers are in separate classes with 200 total classes) and again on the training data with broadly classified species (ex. all Warblers are in the same class with 71 classes total). As shown in Figure 1, specifically classified data achieved accuracy of 54% and the broadly classified data achieved a higher accuracy of 70%, most likely due to the nature of shared attributes between closely related bird species. These accuracies fall into a similar range as current work (50-85% accuracies) [5].

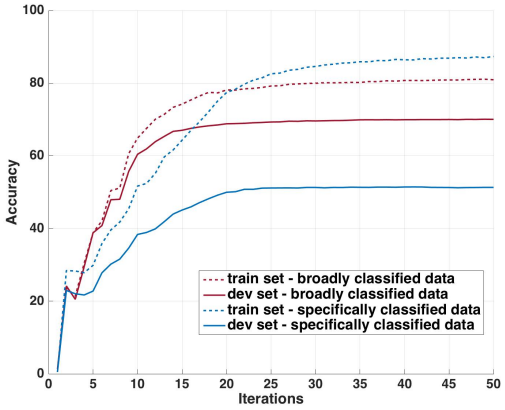

Figure 1: Regularized softmax training and dev set accuracies on broadly and specifically classified species.

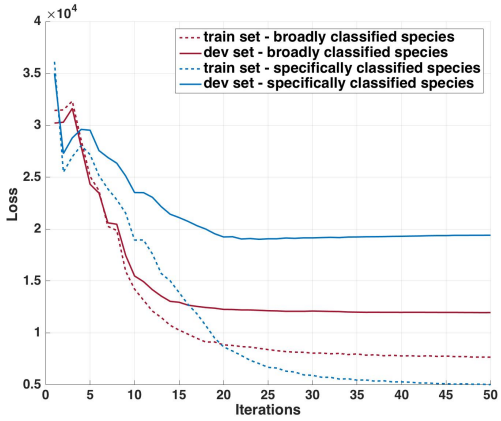

Figure 2: Regularized softmax training and dev set loss on broadly and specifically classified species.

4.2 Multiclass SVM on HOG and RGB Features

The inputs to this method are resized 160x160 RGB images retrieved from the CUB-200-2011 dataset. The feature vectors used in this model are the histogram of oriented gradients (HOG) concatenated with RGB histogram values. HOGs are feature descriptors which have had widespread success in image detection due to its ability to detect silhouettes and invariance to geometric and photometric transforms (except object orientation) (see Figure 4) [6]. Since bird color is essential in identifying its species, RGB histograms of the images are used to aid in identification (see Figure 6). These feature vectors are fed into a multiclass SVM which then outputs its prediction on the species of the bird.

Images are split into the training set and test set recommended by the CUB dataset authors. Using only this method, we obtain a 5% accuracy, which is low perhaps due to the fact that the background causes misinterpretation of the features of the bird (ex. blocked by foliage, reflection over water, etc...). Thus, image preprocessing techniques are needed to separate the background from the bird. Since bird detection in a photo is a project in itself, we utilize the given segmentations and bounding boxes provided by the authors and obtained by humans in the loop. Thus, the process is as follows: original images are cropped to their bounding boxes which are then resized to 160x160 pixels. Then, image preprocessing is done to mask the background using segmentations provided by the authors (see Figure 3). HOG and RGB features are extracted and fed into a multi-class SVM, which boosted the accuracy to 9%. Studies concatenating localized bird head and beak parts and extracting HOG+RGB features to existing features were made with marginal accuracy gains.

Figure 3: Example of image preprocessing to mask background and cropping and resizing of birds using given bounding box data.

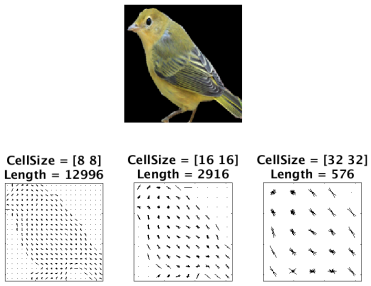

Figure 4: HOG cell sizes were varied, and an optimum cell size of 16x16 was chosen for HOG feature extraction to capture the most informative spatial features of the bird.

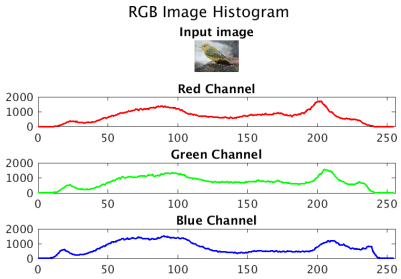

Figure 5: RGB histogram features with background.

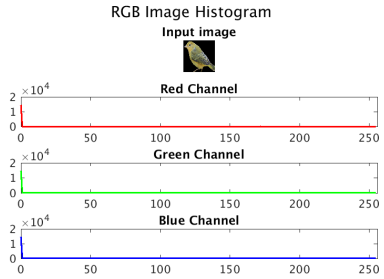

Figure 6: RGB histogram features without background.

4.3 CNNs and Transfer Learning

As an introduction to implementing CNNs, some prototyping is done with the Keras package, a high-level Python deep learning library. A simplified neural network is used (two convolutional layers and 2 fully connected layers) with inputs of 32x32 RGB images from the dataset for fast training. Comparing the byte size of images used in subsection 4.2, the size is already 25 times smaller. Initial results are promising with accuracies of 6% without tuning, so we move on to next step with confidence.

Transfer learning leverages features pre-trained on deep neural networks and allows the user to fine tune a model of their particular interest without having to train the network from scratch [8]. This method uses the pre-trained AlexNet model (5 convolutional layers and 3 fully connected layers) for transfer learning, which has been trained on over a million images from ImageNet [7]. The inputs are 227x227 RGB images from the CUB dataset.

Two methods are used in this study. The first one uses AlexNet as a fixed feature extractor where features are extracted from the last fully connected layer in the network. These features are then fed into a multi-class SVM, and this classifier is used to predict the test set. With this method, we obtain 46% accuracy, which is a large boost to the previous methods. The second method uses fine-tuning, where we import the first 22 layers of the pre-trained AlexNet model and replace the last three layers with a fully connected layer, a softmax layer, and a classification output layer into our 200 species categories. The learning rate (initially set to $1e-4$) in new layers are tuned much larger than the previous layers to speed up the model.

The computational expense between these two methods are large. In one case, we can use a fixed feature extractor method where we have the computational expense of computing the weights of the last layer, which activates 2048 neurons and outputs into 200 classes. In the fine-tuning model, all the weights (including those of the pre-trained network) will be modified from backpropagation, which takes a very long time. We also observe that our CUB dataset is much smaller than the original dataset. Furthermore, ImageNet contains 871 images of birds, so some of the features are already very similar to the CUB-200-2011 dataset. In order to prevent overfitting and to cut down on computational cost, the better method seems to be the fixed feature extractor method.

Figure 7: Mini-batch loss and accuracy of validation training set. After four epochs of training.

5 Summary of Results

Table 1 shows accuracies of the methods described in Section 4. Based on considerations of complexity, computational cost, and accuracy, the best method seems to be pre-trained CNN with fixed feature extraction.

Table 1: Comparison of the accuracies of the different methods.

Method	Accuracy
Regularized Softmax Reg w/ Specific Classes	54%
Regularized Softmax Reg w/ Broad Classes	70%
Image Preprocessing + HOG+RGB and SVM	9%
Pretrained CNN with fixed feature extraction	46%
Pretrained CNN with fine tuning	46%

6 Discussion and Conclusions

In this work, we found that without computer vision we were able to classify birds into broad species categories quite well, but less so with into specific species categories due to similarities in attributes of closely related bird species. We also found that using computer vision on HOG+RGB features of the overall bird gives better results than HOG alone, which makes sense since there is more information and because the colors of the bird are critical in identifying their species. Adding bird head and beak parts as features only gave slight gains in accuracy, possibly due differences in frontal versus side views of the bird. Overall, transfer learning of the pre-trained AlexNet CNN has the most promising results. We found that fine-tuning was slow due to many layers of backpropagation and that using fixed feature gave the most reasonable results in terms of computational cost, accuracy, and preventing over-fitting.

7 Future Work

The next steps that were considering are looking at other algorithms to implement and, more importantly, getting more solid softmax, SVM, and CNN models to gain an understanding of which type of model has better classification accuracy and computational time. Eventually we would like to be able to identify features on the birds with our own algorithm and remove the human classification from the algorithm, moving toward a completely unsupervised learning method. However, this is a huge task and the years of research that have been put into this field, prove just how challenging it could be to get to that point. As discussed, the most promising path forward seems to be the fixed feature CNN model, so in the future we would likely work on developing that model further.

Additionally, one of our original goals had been to track features of a bird in a video. If we were able to develop our CNN further to the point that we could extract specific features from the nodes, we would have a chance at implementing this feature.

8 Contributions

- Anne Alter - worked on the regularized softmax regression models with broadly and specifically classified data; helped develop the CNN model; worked on poster and report
- Karen Wang - worked on the multi-class SVM on HOG and RGB features; helped develop the CNN model and worked on details of running it on cluster computer for faster computing power; worked on poster and report

References

[1] Wah, Catherine, et al. "Multiclass recognition and part localization with humans in the loop." Computer Vision (ICCV), *2011 IEEE International Conference on. IEEE*, 2011.

[2] Berg, Thomas, and Peter N. Belhumeur. "Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[3] Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S., "The Caltech-UCSD Birds-200-2011 Dataset," *Computation and Neural Systems Technical Report, CNS-TR-2011-001*, 2011.

[4] Donahue, Jeff, et al. "Decaf: A deep convolutional activation feature for generic visual recognition." *International conference on machine learning*, 2014.

[5] Branson, Steve, et al. "Bird species categorization using pose normalized deep convolutional nets." *arXiv preprint arXiv:1406.2952*, 2014.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, 2005.

[7] Krizhevsky, A., Sutskever, I., and Hinton, G., "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in neural information processing systems*, 2012.

[8] Razavian, A., Azizpour H., Sullivan, J., and Carlsson, S., "CNN Features off-the-shelf: an Astounding Baseline for Recognition," *Computer Vision and Pattern Recognition*, 2014.