

BIRD CLASSIFICATION

Baomo Zhou, Bowen Deng, Kedi Xia

University of California San Diego, La Jolla, CA 92093-0238

ABSTRACT

The bird classification task is of vital importance because birds play important roles in the ecosystem and are relatively easier to monitor than other species. However, utilizing visual features to evaluate the quantity and diversity of birds is still challenging especially for distinguishing highly similar species. In this paper, we built a random forest model and several CNNs with basenets including vgg16, resnet34, and advanced structure, namely bilinear vgg16 and bilinear resnet34. They achieve classification accuracy of 24.94%, 66.24%, 79.25%, 81.41%, 81.15% on CUB-200-2011 dataset [1] respectively. Our code is available at <https://github.com/den8972/228>.

Index Terms—bird classification, color-related attributes, CNNs, fine-grained classification, bilinear pooling

1. INTRODUCTION

Bird is undoubtedly one of the most important animal groups in the world. Began in the Jurassic Period, birds have been on earth for more than 150 million years and evolved to more than 9,800 species. Besides the extremely huge historical and research values, birds also play very important roles in the current ecosystem by having an enormous variety of applications in real world problems such as environment protection, endangered animal rescue, insect control, pollination assisting, and seeds spreading.

However, with the development of human industry and agriculture, birds are now in danger. Forests are the most important habitats for birds, but they are now disappearing from the earth at a rate of nearly 30 hectares per minute because of the indiscriminate deforestation from us mankind. Besides, the air and water pollution from human factories also brought birds disasters. According to a survey from the Science [2], since 1970s, the North America has lost over 3 billion birds, nearly 30% of the total, and many known birds species, like sparrows and blackbirds are still in decline.

Nowadays, to help ornithologists study and protect birds in a noninvasive way, many computer science researchers are devoted to evaluate the quantity and diversity of birds based on machine learning (ML) and deep learning (DL) methods. The first reason is ML and DL methods have proved to be feasible in many other classification like tasks such as face and

fingerprint recognition. The second reason is birds are relatively easier to monitor than other species and have many distinct features like bills, wings, feathers with different shapes and colors. The last reason is ML and DL methods could sometimes act even better than experienced ornithologists because there are so many bird species and ML or DL methods are born to solve problems associated with big data.

In this paper, the input to bird classification task is birds images of different species. We then use random forest model, CNNs models and improved bilinear models to output predicted species that those birds belong to.

2. RELATED WORK

The current research on bird classification used many machine learning and deep learning methods. Andreia Marini and her colleagues [3] firstly extracted normalized color histogram features and then classified images with support vector machine(SVM). This approach was simple and intuitive but the correct classification rate was around 9%.

With the fast development of GPU, researchers combine handcrafted features with deep neural networks to improve performance. Branson and Steve [4] combined normalized pose feature schemes with state-of-the-art CNNs and largely improved the accuracy to 75%. Similarly, Martin Jaggi [5] split bird songs and noises based on STFT and utilized former ones to train cascaded CNNs. In recent years, some interesting research focuses on fine-grained classification with new basenet like ResNeXt[6]. Besides, Jiongxin Liu [7] introduced a one-vs-most classifier to distinguish between highly similar species of birds. In a different view, [8] changed the basenets structure in front of the classifier by replacing the original pooling layer with a new bilinear one and improved the accuracy to 83%. [9] is a website showing all current fine-grained bird classification results on CUB-200-2011 dataset [1]. The best accuracy of this difficult task was still around 90

Our team noticed that random forest method was seldom used in bird classification. As a result, we built a random forest model in this paper to compare with SVM [3] method. Besides, we also built convolution neural networks based on basenets including vgg16 [10], resnet34 [11] and bilinear models [8].

3. DATASET AND FEATURES

3.1. Dataset

CUB bird 200-2011[1] is selected as the data set for this project, which is a widely used benchmark for fine-grained image classification. The data set is pretty challenging. As can be seen in Figure 1, some species are visually indistinguishable. Even humans need strong domain knowledge to classify them accurately. Bird images are captured in natural scenes, so birds may have different poses, which makes the problem more difficult to solve.



(a) Parakeet Auklet (b) Least Auklet

Fig. 1: Difficult classification example

There are 11,788 images in the data set. Train set has 5,794 images and test set has 5,994 images. All samples are saved in RGB format, and may have different sizes. As many annotations are provided in the data set, it may be used for different tasks. We list some annotations here for reference:

- $\langle image_id \rangle, \langle is_training_image \rangle$
- $\langle class_id \rangle, \langle class_name \rangle$
- $\langle image_id \rangle, \langle class_id \rangle$
- $\langle image_id \rangle, \langle x \rangle, \langle y \rangle, \langle width \rangle, \langle height \rangle$
- $\langle attribute_id \rangle, \langle attribute_name \rangle$
- $\langle image_id \rangle, \langle attribute_id \rangle, \langle is_present \rangle, \langle certainty_id \rangle, \langle time \rangle$

Since bounding box information is provided, the data set can also be used for object detection. Classification is selected as the task because it's an important step in more complex problem like object detection, and we think it's a good start point for us to explore machine learning.

3.2. Features

For random forest, two kinds of features are extracted.

The first one is color histogram which is originally used in [3]. Because color in background is an unfavorable factor, so it's removed first. A pretrained Mask-RCNN[12] is adopted for this purpose. No normalization is required. After that, histograms of each image are calculated. To get a fixed length feature for all images, the number of bins in the histogram is fixed. Also, the number of total pixels after removing background is a variable, so the percentage of pixels in each bin is used rather than the raw pixel number. Color features in RGB space are convenient to calculate, but HSV space has a more close relationship with human perception. We use features from both spaces and compare them.

We utilize attributes provided in the data set as our second feature. Only attributes related to color are kept to make it comparable with the first kind of feature. As attributes are part-based, and each one can have many values, they are more fine-grained than the color histogram.

For deep learning models, features are extracted by CNNs. Features obtained this way are more robust and representative than hand-crafted features. As fully-connected layer, the classifier in the network, requires input to have a fixed dimension, all images are resized to 448×448 . Transformations such as resize, random rotation, random cropping and horizontal flip are performed, in order to improve robustness of the framework.

4. METHODS

4.1. Random Forest

Our first method is random forest, which is inspired by Marini et al[3]. They adopt supporting vector machine for this problem. We think random forest is a better choice. Firstly, random forest fits the multi-class classification nature of the problem. Also, it can handle high dimensional features and large number of training data very well. Last but not least, as shown in Figure 2, it's a kind of ensemble learning method and may help boost the performance.

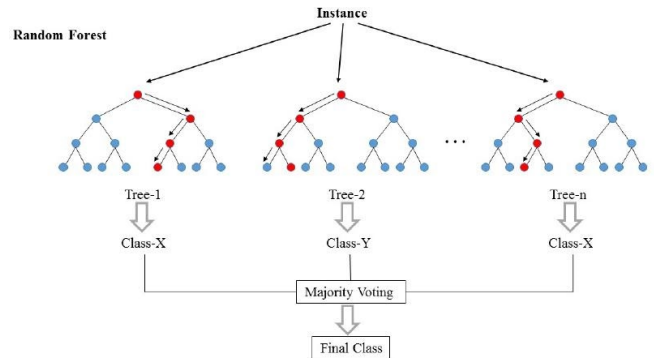


Fig. 2: Random Forest Model[13]

4.2. CNNs

Our second method employs state-of-the-art convolutional neural network models such as VGG[10], ResNet[11]. Furthermore, we perform bilinear optimization on above models to improve their capabilities of addressing the problem of fine-grained classification.

VGG[10] and ResNet[10] are CNN models originally proposed to solve visual problems in computer vision field. VGG is actually a plain and straight forward architecture consisting of basic layers such as convolution, activation, pooling and fully connected layers, but it outperformed many complex models in 2014 ImageNet Large-Scale Visual Recognition Challenge by making improvement over AlexNet through replaying large kernel-size filters with 3×3 filters. Resnet was a more complicated model proposed by Kaiming He and it's based on the concept of residual learning. In other words, every layer is connected to its previous layer via residual connections. Residual connections make it happen that more and more useful features are carried through propagation, then each layer could capture more insightful information than its previous layer by incorporating the residual connection part. When utilizing VGG and ResNet, we perceive the task of bird classification as a vanilla image classification task.

VGG and ResNet already have solid performance in image classification task. However, with regard to fine-grained classification task of which our bird classification task is, a better model capable of capturing subtle difference between subordinate categories is necessary. We show an example of subordinate categories in Figure 3.



Fig. 3: California gulls are subordinate to Ringed-bill gulls

To better wrestle with fine-grained classification, we included bilinear models proposed by Lin et al[8]. As illustrated in Figure 4, bilinear model consists two feature extractors based on CNNs whose outputs are multiplied using the outer product at each location of the image and pooled across locations to obtain an image descriptor. The structure is effective in that it takes advantage of two identical networks as the outer product captures pairwise correlations between the feature channels.

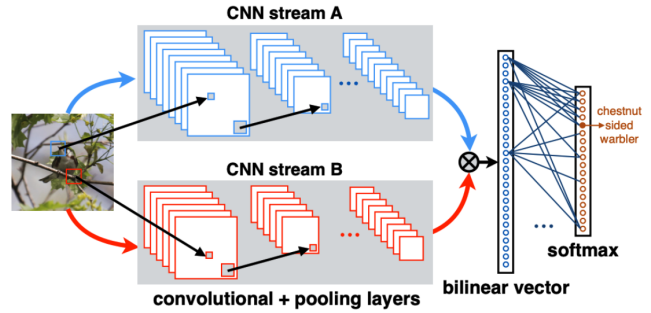


Fig. 4: The structure of Bilinear Model[8]

5. EXPERIMENTS AND RESULTS

5.1. Random Forest

The result we get with random forest is shown in Table 1.

Table 1: Classification Accuracy

Feature \ Model	Random forest	SVM[3]
HSV	10.59%	8.60%
RGB	9.07%	6.86%
color-related attributes	24.94%	/

Random forest trained with color-related attributes has the highest accuracy, which is in line with our expectation. Our result obtained with color histograms outperforms that of [3], but is not satisfying. It demonstrates bird classification is a challenging problem, and is hard to tackle with traditional machine learning methods.

5.2. CNNs

We divided the dataset[9] into two parts with 5994 images for training and 5794 for test respectively. For vanilla CNN models, the training process generally followed the strategy mentioned in paper[10], where the training was carried out by optimising cross entropy loss using mini-batch stochastic gradient descent with momentum. The batch size is set to 8, momentum to 0.9, learning rate to 10^{-4} . The training process takes 50 epochs.

When incorporating bilinear models, we employed a two-step training methodology as described in paper[8]. Firstly, a regular training process was performed to acquire a rudimentary model and achieve 50–60% accuracy with parameters set to values similar to training process above. After 50 epochs of training step 1, a more refined training procedure was carried out where the accuracy of model will be fine tuned to be roughly 80%. Hyper-parameters involved in two-step training process are shown in Table 2. Each step takes 50 epochs to finish.

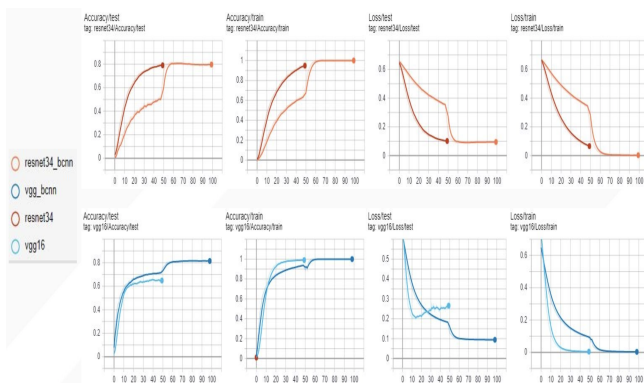
Table 2: Hyper-parameters in two-step training

Step	Parameters	Learning Rate	Momentum	Batch Size
1		0.1	0.9	8
2		0.001	0.9	8

The results we get with CNNs are shown in Table 3 and Figure 5.

Table 3: Classification Accuracy

Basenet	Vgg16	Resnet34	Bilinear Vgg16	Bilinear Resnet34
Accuracy	66.24%	79.25%	81.41%	81.15%

**Fig. 5:** CNNs Results

The bilinear vgg16 model has a large improvement over vgg16 model, which is over 15%. The accuracy of bilinear resnet34 model is close to that of resnet34 model, but still has a 2% improvement. In conclusion, our bilinear models outperform random forest, vgg16 and resnet34 models.

6. CONCLUSION AND FUTURE WORK

We implemented 5 bird classification models including random forest, vgg16, resnet34, bilinear vgg16 and bilinear resnet34. The two bilinear models are the best with an accuracy of around 81% and the random forest model is the worst with an accuracy of around 25%.

We still have many future works to do because we didn't reach the current best accuracy, 83%, of bilinear models. The reason might be overfitting problem. Besides, to get a better result than bilinear models, we could try a better basenet like ResNext [6] or implement a brand new deep learning model like stacked LSTM.

7. CONTRIBUTION

Bowen Deng works on random forest classifier and feature extraction for this part.

Baomo Zhou and Kedi Xia collaborate on the CNN part. They are also responsible for preprocessing the dataset in order to get decent results.

Each team member does literature review for methods related to his part. We collect our results, compare and analyse them together.

8. REFERENCES

- [1] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.
- [2] Elizabeth Pennisi. Three billion north american birds have vanished since 1970, surveys show. <https://www.sciencemag.org/news/2019/09/three-billion-north-american-birds-have-vanished-1970-surveys-show> Accessed Sept. 19, 2019.
- [3] Andréia Marini, Jacques Facon, and Alessandro L Korerich. Bird species classification based on color features. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 4336–4341. IEEE, 2013.
- [4] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets, 2014.
- [5] Elias Sprengel, Martin Jaggi, Yannic Kilcher, and Thomas Hofmann. Audio based bird species identification using deep learning techniques. In *CLEF*, 2016.
- [6] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks, 2016.
- [7] T. Berg, J. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2019–2026, 2014.
- [8] T. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1449–1457, 2015.
- [9] Fine-grained image classification on cub-200-2011. <https://paperswithcode.com/sota/fine-grained-image-classification-on-cub-200/> Accessed June 9, 2020.
- [10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [12] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.
- [13] Casimir. Illustrating the random forest algorithm in tikz. <https://tex.stackexchange.com/questions/503883/illustrating-the-random-forest-algorithm-in-tikz>.

Replies to critical reviews

Disclaimer: We should have received three peer reviews, but we only received two. One of the teams that had to review your project has not submitted a critique. We reported this to TA and have obtained permission from TA to include replies to these two reviews only in the final report.

Critical review from team 48:

- Some images are not the same size.
- Why only chose color features?
- Why bilinear model not Stacked LSTM?
- Why there is no table in result part of CNNs?

Our response to team 48:

- Actually we resized all images into the same size, 448×448 before training.
- Only attributes related to color are used because we would like to compare it with color histograms features we extracted from the image. And we didn't drop any features in CNNs models.
- Bilinear model is much more intuitive than Stacked LSTM and we were short of time.
- We also thought it to be a drawback of the presentation and we would add the table in the final report

Critical review from team 90:

One potential improvement is that when you guys talk about pre-processing, it's not very clear how you removed the background and extracted the features.

Our response to team 90:

The background is removed by a pretrained Mask-RCNN, which is explained in the code demo part. As for feature extraction, it's achieved by counting the number of pixels falling in the predefined intervals with normalization. It might be easier for the audience to understand if we presented some of the code while explaining this part, but limited to the structure of the overall presentation, we failed to do so.