# Bird Image Recognition and Classification

Qihao Zhang
*Stevens Institute of Technology*
qzhang46@stevens.edu

Haosen Yang
*Stevens Institute of Technology*
hyang49@stevens.edu

Bo Su
*Stevens Institute of Technology*
bsu7@stevens.edu

## I. Introduction

The development of mobile Internet, smart phones and social networks has brought massive picture information. According to BI's May article, Instagram uploads about 60 million pictures per day; in February this year, WhatsApp sent 500 million pictures per day; China 'S WeChat Moments are also driven by photo sharing. Pictures that are not restricted by region and language gradually replaced cumbersome and subtle words and became the main medium for conveying words. There are two main reasons why pictures have become the main medium for information exchange on the Internet: First, from the user's habit of reading information, compared to text, pictures can provide users with more vivid, easy to understand, interesting and artistic Information; Second, from the point of view of the source of pictures, smart phones have brought us convenient means of shooting and screenshots, helping us to use pictures to collect and record information more quickly.

But as pictures become the main information carrier on the Internet, problems arise. When the information is recorded in words, we can easily find the content we need through keyword search and edit it arbitrarily. When the information is recorded in pictures, we cannot retrieve the content in the pictures, which affects us from the pictures. The efficiency of finding key content. Pictures bring us a quick way to record and share information, but it reduces our information retrieval efficiency. In this environment, computer image recognition technology is particularly important.

Image recognition is a technology in which computers process, analyze and understand images to identify targets and objects in various modes. The recognition process includes image preprocessing, image segmentation, feature extraction and judgment matching. Simply put, image recognition is how a computer can read the content of a picture like a human. With the help of image recognition technology, we can not only obtain information faster through image search, but also generate a new way of interacting with the outside world, and even make the outside world operate more intelligently. Baidu Li Yanhong mentioned in 2011 that "a new era of image reading has arrived." Now with the continuous progress of image recognition technology, more and more technology companies have begun to involve in the field of image recognition, which marks the official arrival of the era of image reading, and Will lead us into a smarter future.

Bird image recognition and classification play an important role in biology. At present, there are 9,993 species of birds discovered by humans. Image recognition can help humans understand the biodiversity and evolution of birds. Using image classification can use the morphological structure and physiological characteristics of birds to scientifically and reliably identify them. This recognition allows scientists to quickly understand various biological information of the species, as well as its genetic and evolutionary relationships with other birds.

Secondly, bird image recognition can help biologists identify and protect some endangered birds, and it is possible to discover some theoretically extinct birds. Some birds may be too scarce and have not been found in the habitat for a long time. Once they are found, image recognition can be used to quickly identify which species it is, so that they can be protected.

In addition, image recognition can also help discover invasive species to help protect the local ecological environment. If there are invasive species in the area, whether they are brought in accidentally or released by humans, they can be quickly discovered to protect the environment.

In this project, we use Python for modeling and CNN to process images. So as to achieve the purpose of identifying bird images and classifying them.

## II. Algorithm implementation

### A. Dataset Description

The dataset contains 225 different birds in total and it has been found from Kaggle. This dataset includes 31316 training images, 1125 test images (5 for each species) and 1125 verification images (5 for each species). The images obtained by our group are 224 * 224 * 3images in jpg format. It also includes a "merged" image set that combines training, test, and validation images into one dataset. This is useful for our project because we want to create our own training, testing and validation sets. Each group contains 225 sub-categories, one for each species of bird.

The person who provided the data to us on Kaggle clearly stated that the pictures were collected by species name through Internet searches. Once some kinds of image files are downloaded, they use the developed python duplicate image detection program to check if they have duplicate images. Remove all detected duplicates to prevent them from becoming a common image between training, testing and validation sets. Then crop the image so that the grayscale overlaps at least 50 percent of the median value of the image. Finally, adjust the

Fig. 1. Example of COCKATOO

image to 224 * 224 * 3 in jpg format. Downscaling ensures that when processed by CNN, they have enough information in the image to create a highly accurate classifier. All files of each file are also numbered sequentially. Therefore, the test images are named 1.jpg to 5.jpg. It also works for verification images. The training images also use "zeros" to fill the sequence numbers. For example, 001.jpg, 002.jpg... 010.jpg, 011.jpg..... 099.jpg, 100jpg, 102.jpg, etc. When zero is used with the Python file function, the padding preserves the file order and Keras flows from the directory. This saves a lot of time for our project. Each category has at least 100 training image files. This imbalance will not affect my kernel classifier because it achieves an accuracy of over 98 percent on the test set.

Moreover, about 80 percent of the images are male, and 20 percent are female. The color of male birds should be bright, while the color of feathers of female birds is usually cold. Therefore, the male and female images may look completely different. Almost all test and verification images are from males of this species. Therefore, we guess that the classifier may not perform well on female images.

### *B. Algorithm Description*

In deep learning, a convolutional neural network is a class of deep neural networks, most commonly applied to analyzing visual imagery. They have applications in image and video recognition, recommender systems, image classification, medical image analysis, natural language processing, etc.

In this project, we are going to use convolutional neural networks(CNNs) to classify pictures of birds. CNNs use relatively little pre-processing compared to other image classification algorithms.

CNNs are regularized versions of multilayer perceptrons. Multilayer perceptrons usually mean fully connected networks, that is, each neuron in one layer is connected to all neurons in the next layer. The "fully-connectedness" of these networks makes them prone to overfitting data. Typical ways of regularization include adding some form of magnitude measurement of weights to the loss function. CNNs take a different approach towards regularization: they take advantage of the hierarchical pattern in data and assemble more complex patterns using smaller and simpler patterns. Therefore, on the scale of connectedness and complexity, CNNs are on the lower extreme [1].

### *C. Comparison Algorithm Description*

- VggNet. VggNet was born out of the need to reduce the parameters in the convolutional layers and improve on training time.

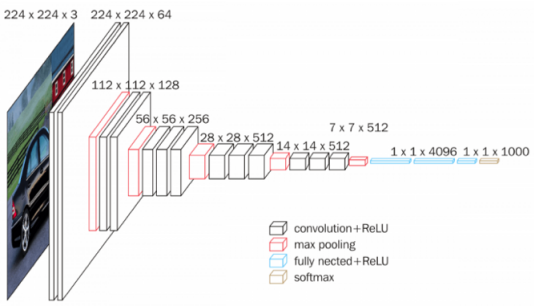


Fig. 2. Vgg Block Diagram

There are multiple variants of VGGNet (Vgg16, Vgg19, etc.) which differ only in the total number of layers in the network. The structural details of a Vgg16 network have been shown below. We use Vgg16 to implement on Birds' dataset.

VGGNet-16 can be divided into 8 parts as a whole, the first 5 segments are convolutional networks, and the last 3 segments are fully connected networks. First, create the first convolutional network. This convolutional network is composed of 2 convolutional layers and 1 maximum pooling layer, which is a total of 3 layers. For the two convolutional layers, the size of the convolution kernel is 3x3, and the number of convolution kernels (the number of output channels) is also 64, the step size is 1x1, and the padding is both 1. This is when we do convolution Several parameters of concern: convolution kernel size, convolution kernel depth, step length, padding. Therefore, the input size of the first convolutional layer is 224x224x3 and the output size is 224x224x64; the input size of the second convolutional layer is 224x224x64, and the output size is 224x224x64; after the two convolutions, there is a 2x2 maximum pooling The step size is 2, so after the maximum pooling layer, the output result size becomes 112x112x64. If padding is used, when kernel size=1, padding=0; when kernel size=3, padding=1; when kernel size=5, padding=2, and so on. That is, we padding only when the size of the convolution kernel is odd.

The structure of the second-stage convolutional network is very similar to the first-stage convolutional network. The size of the two convolution kernels is also 3x3, but the number is 128*128, s=1, p=1. The maximum pooling layer is the same as the maximum pooling layer of the first convolution. That is, 2x2, s=2.

The third convolutional network is different from the first two in that the number of convolutional layers has become 3. The size of the convolution kernel of each convolutional layer is still 3x3, but the number of output channels of each convolutional layer increases to 256, s=1, p=1. The maximum pooling layer is the same as the previous two paragraphs.

The fourth-stage convolutional network is consistent with the third-stage convolutional network. Only the number of output channels of each convolutional layer has become 512.

The last stage of the convolutional network is also 3 convolutional layers with a convolution kernel size of 3x3 plus a maximum pooling layer, but the number of output channels of the convolutional layer is no longer increased, and it continues to be maintained at 512.

- AlexNet. The AlexNet network was designed by Hinton, the winner of the 2012 ImageNet competition, and his student Alex Krizhevsky. After that year, more and deeper neural networks were proposed, such as the excellent vgg, GoogleLeNet. This is quite good for traditional machine learning classification algorithms.AlexNet was born out of the need to improve the results of the ImageNet challenge. This was one of the first Deep convolutional networks to achieve considerable accuracy on the 2012 ImageNet LSVRC-2012 challenge with an accuracy of 84.7 percent as compared to the second-best with an accuracy of 73.8 percent. The idea of spatial correlation in an image frame was explored using convolutional layers and receptive fields.

  AlexNet consists of 5 Convolutional layers and 3 Fully Connected (FC) layers. The activation used is the Rectified Linear Unit (ReLU). The structural details of each layer in the network can be found in the table below. Some advantages of AlexNet: AlexNet uses CNN in a deeper and wider network, and its effect classification accuracy is higher. Compared with the previous LeNet, AlexNet uses ReLU instead of Sigmoid, which can train faster and solve Sigmoid training. The gradient that appears in the deeper network disappears, or the problem of gradient dispersion. In previous CNNs, the average pooling layer was commonly used, and AlexNet all used the maximum pooling layer, which avoided the blurring effect of the average pooling layer, and The step size is smaller than the size of the pooled core, so that there is overlap between the outputs of the pooling layer, which improves the richness of features.
- ResNet. Neural Networks are notorious for not being able to find a simpler mapping when it exists. ResNet addresses this network by introducing two types of 'shortcut connections': Identity shortcut and Projection shortcut. ResNet was proposed by four Chinese including Kaiming He from Microsoft Research. By using Residual Unit to successfully train a 152-layer deep neural network, it won the championship in the ILSVRC 2015 competition, achieving a top5 error rate of 3.57 percent, while the amount of parameters is better than VGGNet Low, the effect is very outstanding. The structure of ResNet can accelerate the training of ultra-deep neural networks extremely quickly, and the accuracy of the model is also greatly improved. ResNet was originally inspired by this problem: when the depth of the neural network is continuously increased, there will be a Degradation problem, that is, the accuracy will first rise and then reach saturation, and then continue to increase the depth will cause the accuracy to decrease. This is not a problem of overfitting, because not only the error on the test set increases, but the error on the training set itself also increases.

  There are multiple versions of ResNetXX architectures where 'XX' denotes the number of layers. The most commonly used ones are ResNet18, ResNet50 and ResNet101. We use ResNet18 to implement on Birds' dataset.

  ResNet18 has around 11 million trainable parameters. It consists of convolution layers with filters of size 3x3 (just like VggNet). Only two pooling layers are used throughout the network one at the beginning and other at the end of the network. Identity connections are between every two convolutional layers. The solid arrows show identity shortcuts where the dimension of the input and output is the same, while the dotted ones present the projection connections where the dimensions differ.

  Soon after the launch of ResNet, Google borrowed the essence of ResNet and proposed Inception V4 and Inception ResNet V2. By fusing these two models, Google achieved an astonishing 3.08 percent error rate on the ILSVRC dataset. It can be seen that the contribution of ResNet and its ideas to the study of convolutional neural networks is indeed very significant and has a strong generalization.

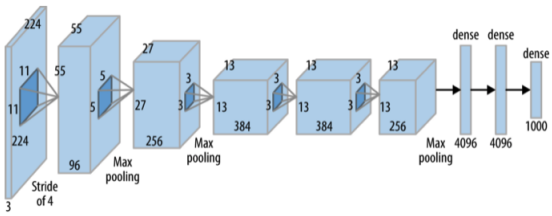


Fig. 3. Alex Block Diagram

### D. Implementation Process

- Load Data. Because the bird images have been sorted into three separate folders: Train, Test and Validation, it

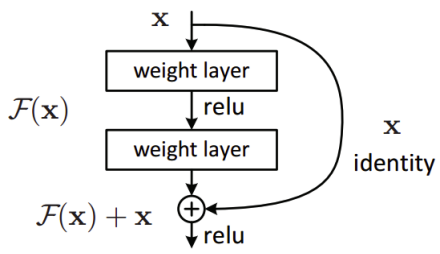


Fig. 4. Residual Block

is convenient to retrieve the image by using Keras built-in function: ImageDataGenerator.

- Build CNN Model. Limited by the PC's CPU computing power and memory size, our initial model adopts a five-layer structure. The first 3 layers are convolutional layers, with 512, 1024, 2048 neuron networks. Using 'ReLU' activation function, 'L2' regularization, and 2*2 Max-Pooling core. Considering the size of the neural network, an 0.2 Dropout was added to prevent the occurrence of overfitting. After the initial operation, there is no overfitting on the validation dataset.
  After 3 convolutional layers, we use flatten function then go into 2 fully connected layers with 512, 225 neurons. Because there are 225 species of birds, the output layer's neuron should be 225 and activation function is 'softmax'.
- Fit Model. We use Categorical Crossentropy as loss function, which is used in multi-class classification tasks and calculates the loss of an example by computing the following sum:

$$Loss = \sum_{i=1}^{outputsize} y_i * log\hat{y_i} \quad (1)$$

  This loss is a very good measure of how distinguishable two discrete probability distributions are from each other. the minus sign ensures that the loss get smaller when the distributions get closer to each other [2].
  As for the optimizer, 'Adam' is used. The 'Adam' is straightforward to implement, is computationally efficient, has little memory requirements, is invariant to diagonal rescaling of the gradients, and is well suited for problems that are large in terms of data and/or parameters [3].

## III. Results

### A. Preliminary Results

After 10 iterations, the accuracy of the training dataset reached 93.27%, and the accuracy of the validation dataset reached 64.62%. Slight overfitting occurred at iteration 5.

The accuracy rate needs to be further improved and the occurrence of overfitting also needs to be solved.

### B. Final Results

The preliminary BirdNet has 3 convolutional layers: 512, 1024, 1024 layers and 2 fully connected layer: 512, 225. we

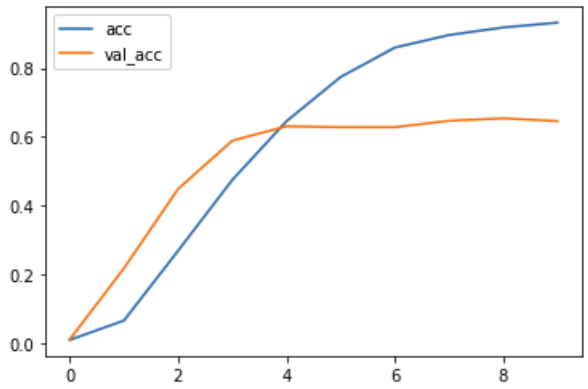


Fig. 5. Train and Validation Accuracy of Preliminary Result

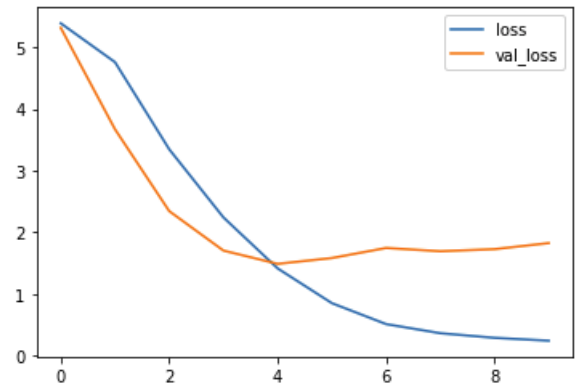


Fig. 6. Train and Validation Loss of Preliminary Result

compared our model with other three models: VggNet, ResNet and AlexNet.

Our dataset contains over 30,000 images of birds, including 225 different species of birds. Although there are many kinds of birds, they are all birds. Different birds can have different colors, different looks, different body types, but they are all birds with two legs and wings, and the head of the bird has roughly the same appearance.

**So the main goal of our network is bird classification, not feature extraction.**

- VggNet has 16 layers, the first four layers only have 64, 64, 128, 128 neurons, the fully connected layer have 512, 512 and 225 neurons, both of which results in the a weak ability to classify. The output accuracy can prove that: after 10 iterations, the accuracy only has 21.47%.
- AlexNet has 8 layers, the convolutional layers have 96, 256, 384, 384 and 256 neurons, but fully connected layers have 2048, 2048 and 225 neurons. Because there are so many neurons in that layers, the bird picture is classified very well: after 10 iterations, the accuracy only has 92.84%.
- ResNet is very different from above two Models. Compared with ordinary networks, ResNet adds a short-circuit mechanism between every two layers, which forms resid-

ual learning. Moreover, the correlation of the gradient actually decays continuously as the number of layers increases. It has been proven that ResNet can effectively reduce the attenuation of this correlation. if we increase the iteration times, ResNet will perform better.

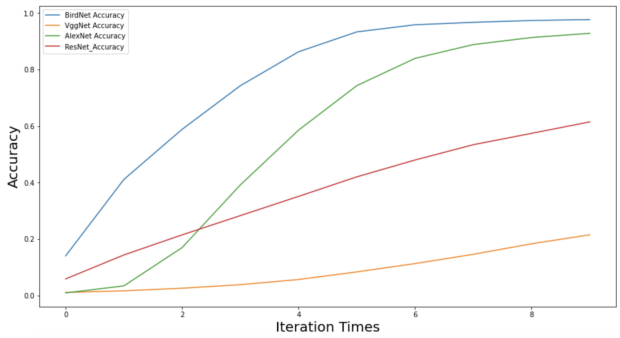


Fig. 7. Comparison Results of the Four Networks

- Final BirdNet. After many comprehensive considerations, we decided to increase the number of neurons in the convolutional layer. Due to the limitations of computer performance, when we tried to increase the number of neurons in the full connected layer, there were too many parameters, leading to the error of resource exhaustion. So, the final BirdNet has 512, 1024 and 2048 convolutional layers and 512, 225 fully connected layers. The final accuracy of BirdNet reached 97.71%.

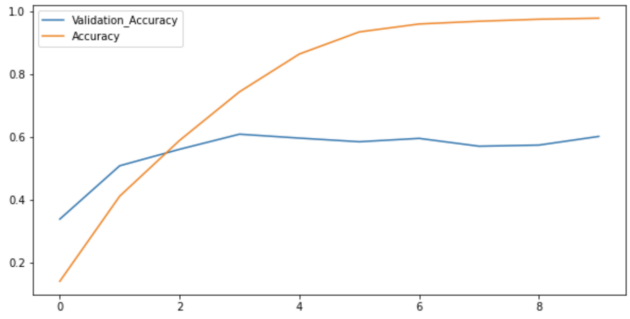


Fig. 8. Final BirdNet Accuracy 1

```
Epoch 10/10
WARNING:tensorflow:multiprocessing can interact badly with TensorFlow, causing nondeterminist
7828/7829 [============================>.] - ETA: 0s - loss: 0.1751 - acc: 0.9772WARNING:tens
7829/7829 [==============================] - ETA: 0s - loss: 0.1752 - acc: 0.9771WARNING:tens
WARNING:tensorflow:multiprocessing can interact badly with TensorFlow, causing nondeterminist
7829/7829 [==============================] - 2223s 284ms/step - loss: 0.1752 - acc: 0.9771
```

Fig. 9. Final BirdNet Accuracy 2

## IV. Further Research Outlook

In the future, deep learning has huge room for development in the field of image recognition. In object recognition and object detection, larger and deeper network structures are tending to be used. Just like the method used in this project, Alex Net in ILSVRC2012 only contains 5 convolutional layers and two fully connected layers. The network structure used by GooLeNet and VGG in ILSVRC2014 both exceeds 20 layers. The deeper network structure makes back propagation more difficult. At the same time, the scale of training data is increasing rapidly. There is an urgent need to study new algorithms and develop new parallel computing systems to more effectively use big data to train larger and deeper models.

In image-related applications, the output predictions of depth models (such as segmentation maps or object detection frames) often have spatial and temporal correlations. Therefore, the study of depth models with structured output is also an important point. Although the purpose of neural networks is to solve general machine learning problems, domain knowledge also plays an important role in the design of deep models. Among the applications related to images and videos, the most successful is the deep convolutional network, which uses the special structure of the image. The two most important operations, convolution and pooling, come from domain knowledge related to images. How to introduce new and effective operations and layers into the depth model through research domain knowledge is of great significance for improving the performance of image recognition. For example, the pooling layer brings local translation invariance, and the deformation pooling layer proposed in this article better describes the geometric deformation of each part of the object. In future research, it can be further extended to achieve rotation invariance and scale invariance.

Studying the relationship between deep models and traditional computer vision systems can not only help us understand the reasons for the success of deep learning, but also inspire new models and training methods. Joint deep learning and multi-stage deep learning are two examples, and more work can be done in this area in the future. The deep learning has achieved great success in practice. The characteristics of the deep model obtained through big data training are eye-catching. The theoretical analysis behind it still needs to be completed in the future. For example, when to converge, how to obtain a better local minima, each layer of transformation obtains those invariants that are useful for recognition, and loses the information, and so on.

Since 2012, deep learning has greatly promoted the research progress of image recognition, which is prominently reflected in ImageNet ILSVRC and face recognition, and it is being quickly extended to various problems related to image recognition. The essence of deep learning is to automatically learn features from big data through multi-layer nonlinear transformation, thereby replacing manually designed features. The deep structure makes it have strong expressive ability and learning ability, especially good at extracting complex global features and context information, which is difficult for shallow models. In an image, various hidden factors are often related in a complex and non-linear way. Deep learning can classify these factors. In the highest hidden layer, different neurons represent different factors. Classification becomes simple.

The deep model is not a black box, it is closely related to the traditional computer vision system, but it allows the

various modules of the system (that is, the various layers of the neural network) to be jointly learned and optimized as a whole, thereby greatly improving performance. Various applications related to image recognition are also promoting the rapid development of deep learning in various aspects of network structure, layer design and training methods. We can foresee that in the next few years, deep learning will enter a period of rapid development in theories, algorithms, and applications. We look forward to more and more exciting work that will have a profound impact on the academic and industrial circles. The great success of deep learning in image recognition is bound to have a significant impact on various multimedia-related applications. We look forward to more scholars in the near future to study how to use the image features obtained by deep learning to promote the rapid progress of various applications.

## V. Conclusion

Bird images means a lot in biology, identifying and protecting birds in wild.

Our initial experimental result is that the accuracy of the training data set is 93.27%, and the accuracy of the verification data set is 64.62%. The accuracy is low and there is a slight overfitting. After that, we compared our model with the three models VggNet, ResNet and AlexNet. According to the excellent performance of ResNet, for example, it can effectively slow down the attenuation and have better performance when the iteration time is extended. We increased the number of neurons in the convolutional layer. The final accuracy of BirdNet has been improved to 97.71%.

Through experiments, we found out what efficient convolutional neural network has in image-related applications. This can help us classify birds efficiently and accurately. This shows that in addition to using convolutional neural networks in bird image classification, we can also extend to other fields, such as face recognition and so on.

In addition, deep learning also plays an important role in image recognition. Its deep structure makes it have strong expression and learning ability, especially good at extracting complex global features and context information, which is difficult for shallow models. In the image, various hidden factors are usually related in a complex and non-linear way. Deep learning can classify these factors. In the highest hidden layer, different neurons represent different factors. Classification becomes simple.

In the future, convolutional neural network has greater potential not only to recognize pictures and videos, but also to imitate. Imitate the picture/video style. we are willing to do that.

## References


[1] Wikipedia contributors. "Convolutional neural network." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 12 Nov. 2020. Web. 17 Nov. 2020.

[2] https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/loss-functions/categorical-

[3] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[4] https://towardsdatascience.com/the-w3h-of-alexnet-vggnet-resnet-and-inception-7baaaecccc96