
Transfer Learning for Dog Breed Identification

Yizhou Chen
yic244@ucsd.edu

Zhongke Ma
z2ma@ucsd.edu

Qichao Zheng
q5zheng@ucsd.edu

Yisheng Ji
y3ji244@ucsd.edu

Abstract

In this project, we achieve classification of dog breeds from Stanford Dog Dataset which contains images of 120 dog breeds. ResNet50, DenseNet, VGG19 and InceptionV3 were used to extract the features. To get a better accuracy, we apply transfer learning on different networks such as VGG19 and InceptionV3. These networks were already built on Keras, which is a deep learning library and provides some popular neural networks. The result of this project can be used in identification of dogs in images.

1 Introduction

1.1 Motivation

Dog have become more and more important in our daily life, they appear as pets, transportation, drug detector and so on. Since different breed of dogs can help people in different way, the identification on breed of dogs have become commonly mentioned nowadays in many fields. Thus, we need a more accurate and efficiency way to classify the breed of dogs.

1.2 Convolutional Neural Network

Deep learning models have achieved remarkable results in computer vision in recent years. [1] As one of the deep learning networks, Convolutional Neural Network has widely used around the world for image classification. The CNN consist of neurons and the weights and biases of the neurons can be trained. Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linearity. Convolutional layer, pooling layer, and fully-connected layer are the three main types of layers of Convolutional Neural Network.

In order to improve the performance of the network, CNN model nowadays has become deeper. However, as CNN become deeper, the cost of time and computing resource also become larger. Thus, to save time and make our calculation more efficient, we will use a pre-trained model in our project. Pre-trained model is a model already trained for some other task, we can re-train them in order to fit our task.

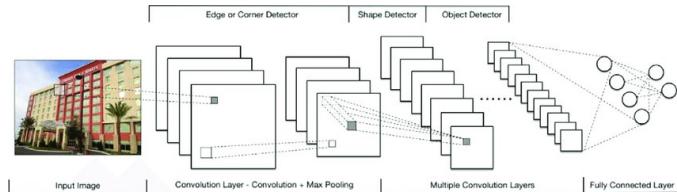


Figure 1: Basic structure of CNN.

2 Description of methods

In this project, we will adopt several CNN models for the purpose of image classification. The past ten years have witnessed great success in the performance of deep network handling the problem of image classification. And there have been several great framework or architectures successfully achieve below the 5% top five error rate. Such as the **DenseNet**, **ResNet** and **Inception**. The former two methods adopt the concepts of *shortcut connections*, while the latter one utilize the manner of *factorizing convolutions*. The purpose of these methods is intuitive, by reducing the parameter of deep network so as to allow for the network to grow deeper and deeper.

Therefore, in this project, we wish to explore the difference between these methods. Namely, we wish to see the difference of *a) shortcut connections* and *b) factorizing convolutions* in performing large scale, multi-labeled image classifications. Unlike the basic classification problems, where the objects being classified varied in a wide range. In this project, we wish to examine how well the deep network can handle the classification of similar objects.

VGG19, DenseNet, InceptionV3 and transfer learning were used to build our neural networks.

2.1 VGG19

2.1.1 Introduction

VGG is a network of increasing depth using an architecture with very small convolution filters. With small convolution filters, the number of parameters can be reduced. The input of VGG is 224*224 RGB images. Then the input image will go through the convolutional layers, where filters with a small receptive field will be applied. Each convolutional layer has different depth, some of the convolutional layers will have a max-pooling layer followed, and the max-pooling layer has a 2×2 pixels window with stride 2. With max-pooling layers, VGG can achieve spatial pooling. After the stack of convolutional layers three fully connected layers are followed. The first two layers have 4096 channels each, the third have 1000 channels. And the last layer of the network is a soft max layer. After the input goes through the whole network, the output will be classification contains 1000 channels. VGG16 and VGG19 are the two most common models of VGG. VGG 16 has 16 weight layers which has about 138 million of parameters in total, while VGG19 has 19 weight layers which has about 144 million of parameters in total. [4] The model we will use in our project is VGG19, it can achieve 7.3% top 5 error in ILSVRC.

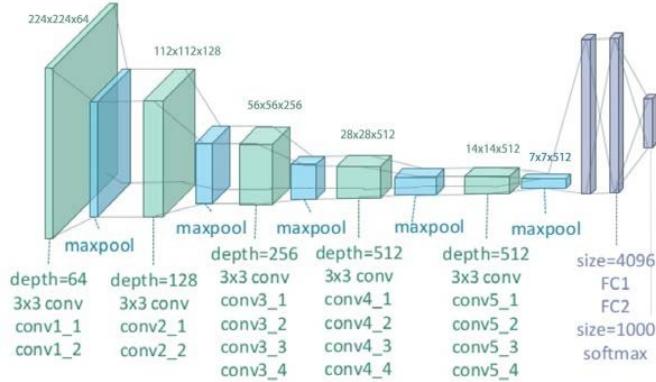


Figure 2: VGG19 architecture.

2.1.2 title

2.2 DenseNet

More and more models have shown that convolutional neural networks can become substantially deeper, more accurate, and efficient to train if they contain shorter connections between layers close to the input and those close to the output. Thus, Dense Convolutional Networks (DenseNet) was introduced. DenseNet consists of dense block, within each dense block, each layer is connected to

every other layer in a feed-forward fashion. [3] In DenseNet, every layer is connected together by dense connectivity in order to have max information flow between the layers which means every layer have multiple inputs from other layers.

In DenseNet, fewer parameters is needed, which means the parameters in DenseNet can be more efficiency than traditional neural networks. Besides, DenseNet can be deep while the layers can be slim. With all these advantages, DenseNet is easily to be trained.

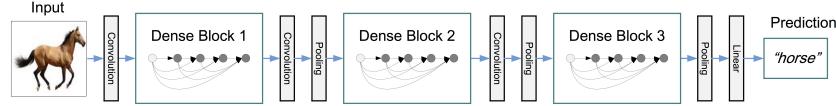


Figure 3: A deep DenseNet with three dense blocks.

2.3 InceptionV3

Inception is a deep convolutional neural network which provides a new state of the art for classification and detection. It makes a better use of the computing resources inside the network and makes the network deeper and wider while the computing budget is not affected. Inception networks consist of Inception modules. The main idea of the Inception architecture is to consider how an optimal local sparse structure of a convolutional vision network can be approximated and covered by readily available dense components. [5] To find the optimal local sparse structure, we assume that each unit of the previous layer refers to some part of the input image, which can find the clusters that focused on some particular region, thus, these clusters can be summed up by a 1×1 convolution in the following layer. And to avoid patch-alignment problem, the filter size of Inception has only 3 options: 1×1 , 3×3 and 5×5 . To keep the computing complexity stable, Inception network uses dimensionality reduction which can reduce the feature maps from 192 to 16.

InceptionV3, as the improved model of Inception networks, has 42 layers, which is much deeper than GoogLeNet, the first model of Inception networks. The improvement InceptionV3 make is that they applied factorization of convolutions and improved normalization in the networks. [6] Comparing with GoogLeNet, the top 5 error of InceptionV3 is improved from 6.67% to 4.49%.

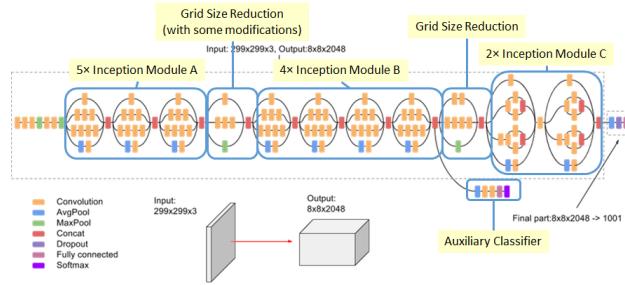


Figure 4: InceptionV3 architecture.

2.4 Transfer learning

Transfer learning is a machine learning technique where a model trained on one task is re-purposed on a second related task. In traditional CNN for image classification, usually the earlier layers are to detect edges, while deeper layers will try to detect some specific features. With transfer learning, we can use the earlier layers to re-train the deeper layers. The main benefit transfer learning is saving training time without hurting the performance. Also, with transfer learning, we can use less data to train the neural networks.

3 Implementation Details

In this section, we

4 Experimental Settings

4.1 Datasets

The dataset we used in this project is Stanford Dogs Dataset, which contains images of 120 breeds of dogs from around the world, such as Chihuahua, Japanese spaniel and Rhodesian ridgeback. This dataset has been built using images and annotation from ImageNet for the task of fine-grained image categorization. There are totally 20580 images in this dataset, 12000 images in the training set and 8580 in the testing set.



Figure 5: Stanford Dogs Dataset.

4.2 Training Parameters

4.3 Hardware Used

5 Results

5.1 InceptionV3

5.2 DenseNet121

5.3 VGG19

Number of selected categories	Accuracy
20	87.31
40	83.53
60	80.01
80	76.29
100	74.42
120	70.67

Table 1: Accuracy versus categories selected

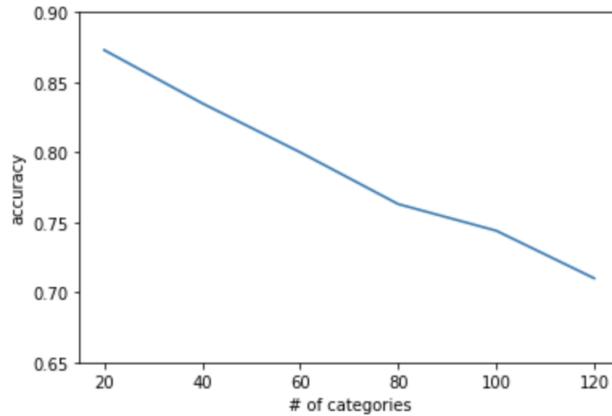


Figure 6: Accuracy versus categories selected

The test accuracy of our algorithm is 70.07% if all 120 dog breeds are included.

5.4 Comparision

6 Discussion



Figure 7: Some correctly classified samples



Figure 8: Some misclassified samples

From the above misclassified samples, the color of the background is so much close to the dog, which will We also notice that in the training test there exist some images contains more than one dog and same image appears twice in two different breeds as shown in below. This type of error seems not solvable via programming.



Figure 9: Sample of more than one dogs

7 Appendix

References

- [1] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.
- [2] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
- [3] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//CVPR. 2017, 1(2): 3.
- [4] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
- [5] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.
- [6] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.