COMP576 Project Proposal
Multi-class Dog Breed Classification

Oybek Khakimjanov (ojk1), Wenbo Zhao (wz47), Kaiwen Deng (kd45)

Nov. 2022

1 Introduction

In this project, we would like to use Stanford University Dogs Dataset [2] to train multiple deep learning models and our own multi-layer model. The trained models should be able to: accept a picture of a dog and determine the breed of the dog.

Then, we will compare the performance of different models by a group of pictures of different breeds of dogs as test sample. In addition, we will visualize the training results to make the comparison between the different models more visual, and then analyze the advantages and disadvantages of different models in the application of dog breed classification.

2 Design and Specification

For this project we will be using **Stanford Dogs** dataset presented by Stanford Vision and Learning Lab.

This dataset provides images of **120 breeds** of dogs from around the world. The contents of the dataset are the following:

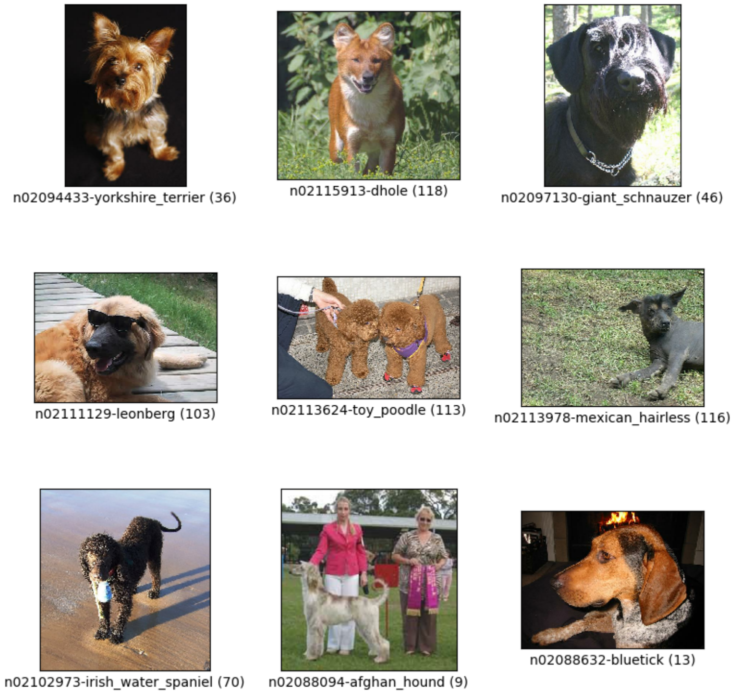
- Number of categories: 120
- Total number of images: 20,580
- Annotations: Class labels, Bounding boxes
- Dimensions: Only images of 200 x 200 pixels or larger are kept in the dataset

The dataset will be split into training (65%), testing(30%) and validation(5%) sets. In total, this dataset represents the most comprehensive and diverse dataset for breed classification. Each image is annotated with an object class label see Figure 1.

This dataset is particularly challenging due to a variety of reasons. First, being a fine-grained classification problem, there is little inter-class variation.

For example, certain breeds of dogs may be very akin to one another and share similar characteristics and colors. Second, most of the images in a dataset contains humans and other unrelated objects, thus, increasing the noise caused by background variation. Unlike other fine-grained visual dataset, this dataset has larger number of images per category, and thus it can be used for accurate testing of different learning and optimzation algorithms.

Figure 1: Sample from Dogs Breed Dataset



n02094433-yorkshire_terrier (36)

n02115913-dhole (118)

n02097130-giant_schnauzer (46)

n02111129-leonberg (103)

n02113624-toy_poodle (113)

n02113978-mexican_hairless (116)

n02102973-irish_water_spaniel (70)

n02088094-afghan_hound (9)

n02088632-bluetick (13)

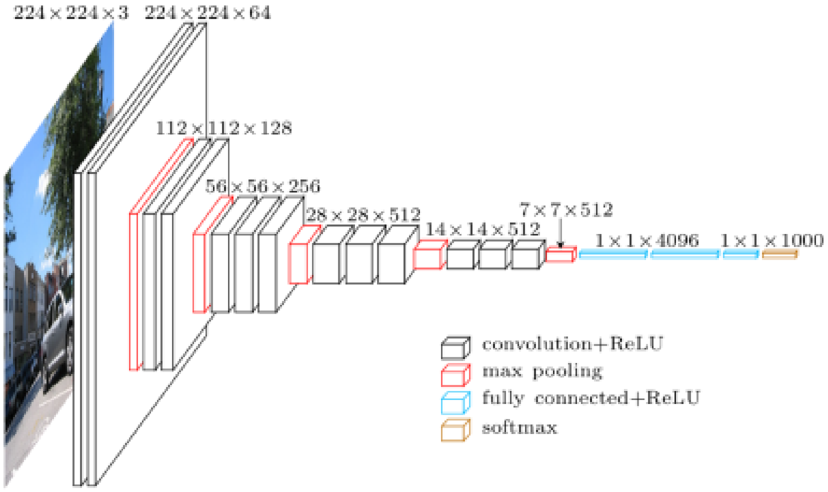
3 Model

In order to do benchmarking, we plan to compare several Computer Vision models plus our own model in this project.

3.1 VGG-16

The VGG-16 [4] model is one of the most popular pre-trained models for image classification tasks. Brought up in 2014, it beat the top-performance model AlexNet [3] and still keeps high-level performances today. This model is sequential in nature and uses a lot of 3×3 filters. At each stage, these small filters are used to reduce the number of parameters followed by the ReLU activation function. Compared to AlexNet, VGG-16 is a much larger model with a huge amount of parameters, which makes it much slower when training. But in general, this model is simple, intuitive, and one of the top players in image classification tasks. The architecture of VGG-16 is shown in Figure 3.

Figure 2: VGG-16 Model



3.2 Resnet

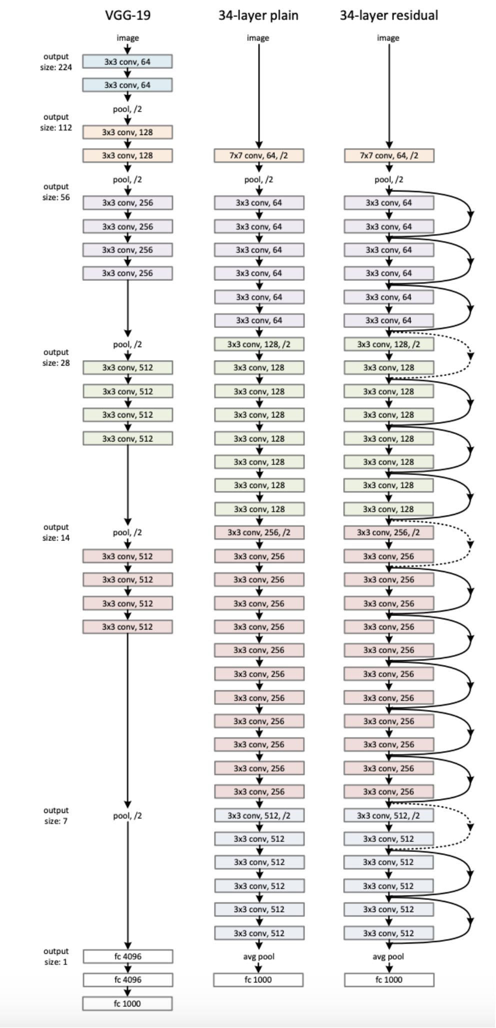
The Resnet152(Residual Net) [1] is not the first model of Resnet but it offers great performance on image classification tasks.

There was a major problem for the deep learning model before Resnet: as the model went on to become deeper, the accuracy became poorer. So this problem became the main motivation of the Resnet model. Surprisingly, the Resnet model can also tackle the Vanishing Gradient issue.

In the Resnet model, after a 7×7 convolutional layer and a max-pooling layer, there are 4 similar layers with different filter sizes, but all of them use 3×3 operation. The architecture of Resnet is shown in Figure 4.

However, the key point in this model is: after every two convolution layers, the model bypasses the layer in between. These skipped connections are called ”identity shortcut connections”. As a result of using this, the model is able

Figure 3: VGG-16 vs Resnet



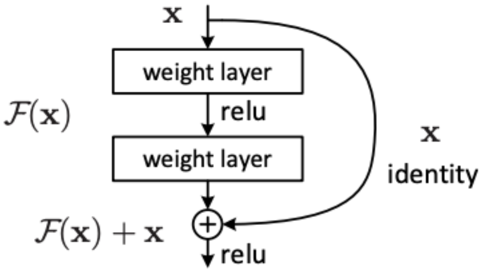
to pass the original information to the deep convolutional layers. The building block of Resnet is shown in Figure 5.

3.3 Inception

The Inception model [5], as known as GoogLenet, was another astonishing image classification model in 2014. Comparing to VGG and AlexNet, it is much smaller in the size with only 7 million parameters and relative low error rate.

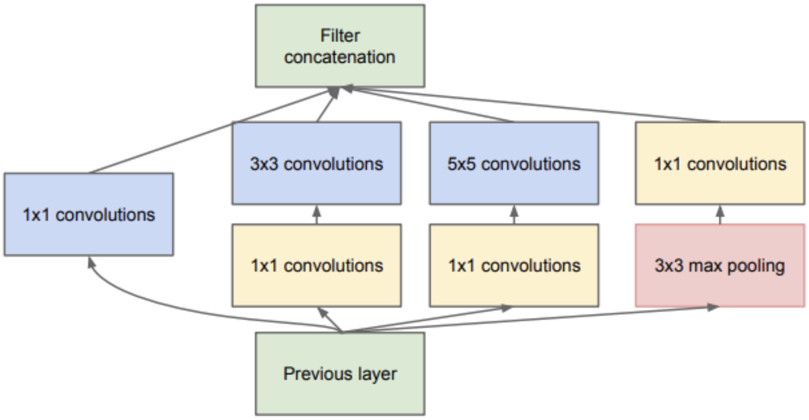
The fundamental part of the Inception model is the Inception Module. Simply speaking, this module performs convolution with different filter sizes or max

Figure 4: A block in Resnet


x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

pooling on the input, then concatenates the result for the next Inception Module. The Inception Module is shown in Figure 6.

Figure 5: The Inception Module

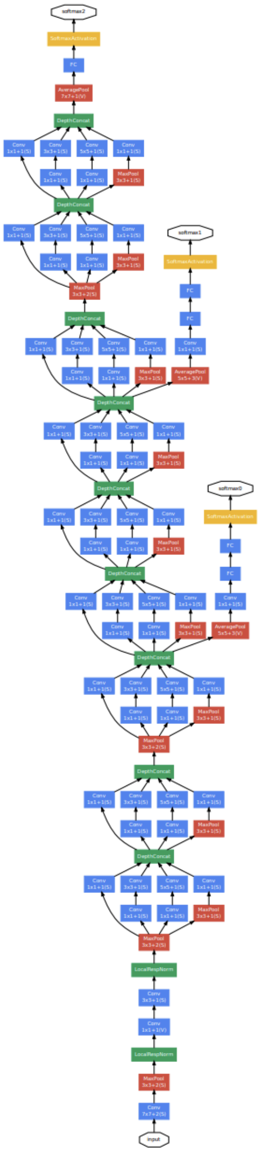


With that design, the Inception model can achieve astonishing results with only 22 layers. And the newest InceptionV3 introduces the batch normalisation and RMSProp optimizer to the model, and adds more factorization, which largely increased the accuracy and made the model less complex. The architecture of Inception is shown in Figure 7.

3.4 EfficientNet

The EfficientNet model [6] proposed by Google is another major progress in the Computer Vision field. In this paper, they proposed the Compound Scaling - a new Scaling method which proposes that if we scale the dimensions by a fixed number and do uniformly, the model will achieve much better performance. The architecture of EfficientNet is shown in Figure 8.

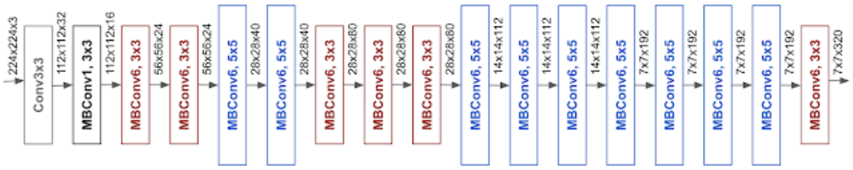
Figure 6: The Inception Model



3.5 Our Model

Although we haven't decided on the details of our model, but the general idea is that we use existing pre-trained model as our base model, and layer them up and combine them with max-pooling, dropout and relu layers. Therefore, in our report we can compare all these pre-trained models' performance with

Figure 7: The EfficientNet Model



our model to see whether complex layered-up model can be better than these models.

4 Potential Issues

While interesting, image classification for this dataset can be a very computing-demand task for a machine. Several challenges particular to the image classification problem have been noted by us.

Firstly, as the dataset(Stanford University Dogs Dataset) we are preparing to use is a large-scale dataset and the free memory of the Google Colab free version is limited, we may come accross conditions when we overpass the limitation, which will hinder the process of our project.

Secondly, there are only slight differences between some dog breeds. So it might be difficult for the model to properly classify the dog breeds. We need a lot of research to see how to handle this problem properly.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[6] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.