# Summary for Assignment_2

Yanxi Li, yli130@kent.edu

## Conclusion:

For this assignment, increasing the training sample size for each classification can improve the model's prediction accuracy, but not that efficient. From steps 1-3, the best performance based on my result is 0.904 which the training sample has reached 5000 for each classification.

However, when it comes to pretrained neural networks, the testing set performance is 0.979 even with the step1 small training sample size. And increasing the training sample size in steps 2-3 still shows similar results. The pretrained model shows a powerful result because it was previously trained on a large dataset that does not only include cats and dogs. So, the model performance mostly relies on the pretrained model feature extraction instead of the training dataset sample.

In this way, the pretrained neural network used in image processing is more important than increasing the training dataset. It should be noted that the pretrained model should be related to the current classification, if the pretrained model has a huge difference from our classification, the result could be worse.

## The following are details in each question:

### Basic model without any method to reduce overfit:

The basic model I created from scratch without any method to reduce the overfitting, the accuracy is only 0.683. From the Training and validation plot, the training accuracy keeps increasing until reaches almost 100% whereas the validation accuracy only reaches around 70%, which indicates that overfitting happens.

### Question_1:

First, I added the dropout since it shows improvement in reducing the overfit in the densely connected layers. The testing set accuracy was improved to 0.714 but still overfitting.

Data augmentation was added on the dropout basis because it's a powerful technique to reduce overfit for image processing. After the data augmentation, the model accuracy was increased to 0.808 which looks great, and the overfitting problem was reduced according to the training accuracy plot.

## Question_2:

I increased each classification training sample from 1000 to 3000 and retrained the model, 2 testing set accuracies are 0.797 and 0.898 which are predicted for the model without and with data augmentation, respectively. In addition, model without data augmentation also has the overfitting problem. We can draw a basic conclusion that increasing the training data sample size with data augmentation can improve the model performance effectively.

## Question_3:

I decided to keep increasing the training sample size for each classification, from 3000 to 5000. This time model without augmentation and with augmentation accuracy are 0.843 and 0.904. Again, the model without augmentation still has the overfit problem. The model with data augmentation has improved above 90%.

Although the model prediction has reached 0.904 this time, for the appropriate prediction model, I still want to choose the 3000 training samples for each classification. The reason is as follows, the model was improved a little from 0.898 to 0.904 but at the same time, our training sample is 10000 in total instead of 6000. If the computational resource is not enough, increasing the training sample size to 10000 is not an economic method.

## Question_4:

This step comes to the pretrained VGG16 model. I first generate the model with fast feature extraction without data augmentation with the small data in Question_1, the validation accuracy reached to around 97% this time but the model still shows overfit from the figure. Then with augmentation, we need to freeze the convolutional base which is computationally expensive in contrast with the last model. With augmentation, the model prediction is 0.977, which is a strong result compared to the previous model from scratch. The last step is fine-tuning, which consists of unfreezing a few top layers of a frozen model base used for feature extraction. Here I only unfreeze the fourth layer which is the top layer. As a result, fine-tuning pushes the prediction performance even further to 0.979.

When I increase the training sample size following steps 2-3, the result shows only a little improvement (from 0.979 to 0.982) since 0.98 is pretty high.