# **Machine Learning Through Cats and Dogs Classifier**

**Anuradha Tiru-Narayanan, Gavyn Gallagher, Jhoseline Vasquez, and Venkat Rajagopal**

**Northwestern University**

**MSDS 422: Practical Machine Learning**

## **Introduction**

Our company, XYZ, allocated our team to evaluate a dataset consisting of images of various cats and dogs. Our goal is to accurately identify whether the image contains a cat or a dog through machine learning techniques. The training dataset consists of 25000 images while the test dataset holds 12,500 images.

**CNN model #1**

CNN uses max pooling to replace output with a max summary to reduce data size and processing time. In this model, activation method is relu, max pooling size is 5. Softmax is applied to the last layer. Regression with loss as 'categorical\_crossentropy' is used and the optimizer used is adam. Accuracy of the training with loss of 0.4551 data is 0.78880.

**CNN Model #2**

Developed model 2 using SGD optimizer, relu, he\_uniform initializer, as the parameters with three blocks of 32, 64 and 128 and also categorical\_crossentropy and learning rate .001. Achieved the best accuracy of .9866 with the training data set and loss rate of .1723. Running the prediction of the test data set produced the loss rate of 0.1405012458562851 and test accuracy of 0.9710000157356262. Achieved kaggle score of .14989 on the test data. Following are excerpts from model summary: Total params: 22,961,706, Trainable params: 10,003,490 and Non-trainable params: 12,958,216.

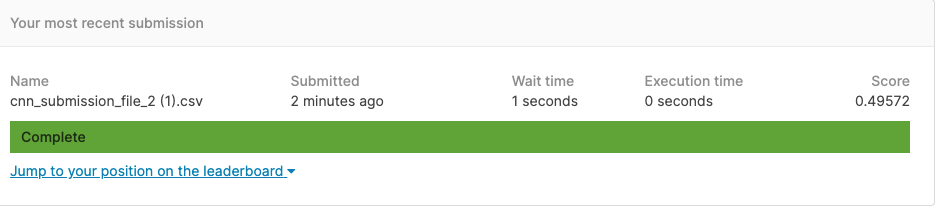
**CNN Model #3**

We tried two variations in model 3. One with a training generator with the batch size is 100, steps\_per\_epoch as 250. This model is to use the complete training data set around 25000 and 1000 images for validation data. It ran for about 2.5 hours while fitting the model but showed the training and validation accuracy/loss as a straight line. The other model has around 2000 images for training data and 1000 images for validation data. The test directory has about 12500 photos. The second model performed better than the first one. Calculate the loss on training and validation and interpret how well the model is doing for these two sets. The blue line represents the accuracy of trained data. The yellow line represents the accuracy of the validation dataset. Accuracy of the training data set went up for every epoch, the accuracy of the validation data set almost remained the same with slight variance. Accuracy of training loss went down per epoch; validation loss went up. Training accuracy of the model is 0.9920 with loss of .0224.

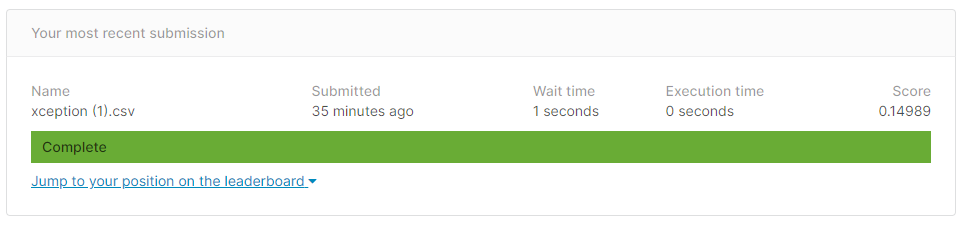
**Conclusion**

Based on the submission responses from Kaggle it is determined that model #2 produced the best results.**Appendix**

CNN



CNN-2



CNN-3

