To: Wendy Fisher
From: Manisha Jaiswal, Taylor Kramer, John Santiago
Date: 4/30/2020
Subject: Project Update Part 1

Pupper Lovers!!!

We are Group 12 and we decided to be the *Pupper Lovers* Team! Our Team members are Manisha Jaiswal, Taylor Kramer, and John Santiago. Through discussion in what we wanted to do for our topic, we decided to do something related with dogs since we all love dogs! That is when we thought of a Dog Breed Image Identifier because how often have you seen a picture of a very adorable puppy scrolling through social media, but you cannot identify the breed of the dog? For most people especially those who are not dog experts, it is difficult to simply identify any dog's breed. Therefore, our proposed machine learning solution would be to create a Dog Breed Image Identifier that can take in the picture of the dog and tell you what type of breed the dog is!

Through further research we discovered that something like this has been implemented before; we found a link of dog breed classifier that uses AWS Sagemaker [1]. For our proposed machine learning solution, we are currently experimenting with a convolutional neural network (CNN) and the libraries of TensorFlow and Scikit-Learn to create it. The trained model takes in the image of a dog that is converted into the form of a tensor or array that can then be identified as one of the 120 different dog breeds in the training dataset and ideally the correct dog breed gets outputted.

In determining what data, we wanted to use, we decided to use the Stanford Dogs Database which consisted of 120 different kinds of dog breeds and 20580 images in total. In the link to the project using AWS Sagemaker, they had already provided two files for this data: one was a list of links for the 16,464 photos used in training and the other was for a file with a list of 4,116 photos used in testing. This helped immensely in creating our training and testing sets

which are an 80% 20% split as there was no need to clean/wrangle the data. Because this data was provided clean to us with good quality photos, we decided to use the two files we found to call in each photo straight from the internet, one at a time, turn them into matrices, and then insert the data into a NumPy array ready to be fed into the model.

The process of creating the dataset does take about 50 minutes with so many images to process but once this is complete everything is set to move on to the model training portion of this project. In our case, feature extraction is not a necessary step because the images already contain all the information, we want to train the model on. It would probably be less helpful for us to try and pull out any other features that might help identify the dog breeds because there is the chance that we would oversimplify the data at that point. So, from here we just move on to building our model.

Moving on to our model, the features we used were 120 different types of dog breeds within the training and test set. After converting each dog image into an array to be fed into the model, we were able to create our first preliminary model using a convolution neural network. While we got accuracies in the 80s when training our model, when evaluating the model using our test data, we achieved a very low accuracy of 5.1 percent. Note that for these beginning stages we have only been using about 0.2 percent of the image data to save time and for now we do seem to be dealing with issues of overfitting possibly because of not using enough data in training the model. While our first attempts have yielded very low accuracies, we have promising plans of upgrading the model using several different techniques.

Last part of the project we have going so far that is worth mentioning is how we are testing our model. For us, we think it is important to judge our model on accuracy first and for most. But we wanted to take it one step further by finding a picture the model has never seen

before and see if the model can identify the dog in the photo. We chose to find a picture of a Siberian husky, one of the dog breeds in the set, and use model.predict() and see if our model is able to output the correct breed. The team agrees that if the model can successfully complete this task, then it serves its purpose and reaches the goal we are aiming for.

One constraint and challenge we are currently facing is that creating our data set takes close to approximately 50 – 60 minutes to run. To mitigate this problem for now we decided to use only about 20 percent of the data to first ensure that our code is running and create a somewhat promising model from that data set before using all the images. Another constraint we are encountering is figuring out how to increase the accuracy of our model. Through further research, we concluded that transfer learning could provide us with a huge boost in our accuracy and model effectiveness. By taking in features that can reach "far deeper" than our model could in a reasonable amount of time then by adding a few layers to these transferred features so that our model complements our dataset then we might be able to reach accuracies that would be unattainable otherwise.

Our second idea for improving our model is by improving our dataset through data augmentation. Given that there is a lot happening in any kind of picture of a dog, by augmenting our data we not only increase the amount of data we can train with, but we also allow our model to see each photo from different angles and in different lighting making it more adaptable and robust.

For our project expectations, we aim to improve the accuracy of our model to be about 90+ percent. We do not expect to hit these values until after implementation of transfer learning and data augmentation. From there we intend to make micro adjustments to our layers and dropouts to avoid any sort of overfitting while increasing our accuracy as much as possible. If we

get to the point where we can identify dog breeds in random pictures off the internet, then we will consider the job complete.

Lastly, a brief look at our timeline. Our goal was to have a model, any model, up and running at this point. As can be seen from our discussion, we have successfully accomplished this and more. We are done with creating our datasets, we have created a running model, and we have already moved on to the next portion of the project where we are attempting to adjust and improve our model. Our next goal is to have a model that is providing us with high, 90+, accuracies by 5/08 and so far, we are on the right track to obtaining that goal.

- *Timeline for remainder of the semester:*

05/02/2021 – Update – Report Submittal / Peer Reviews of Reports. (Manisha, Taylor, John)

05/08/2021 – Model is predicting dog breeds with high accuracy. (Manisha, Taylor, John)

Week of 05/09/2021 – Working of final video. (Manisha, Taylor, John)

05/13/2021 – Final Video – Video Submittal / Peer Review Videos. (Manisha, Taylor, John)

- *Overview of Team member task assignments:*

Taylor – preprocessing data and creating first model outline.

Manisha – deep dive researching model parameters.

John – Start looking into how we are going to best report our results.

(But ideally, we all are working together so that everyone can learn from each other.)

References

[1] J. Burt, “Build Dog Breeds Classifier Step By Step with AWS Sagemaker,” *KDnuggets*. [Online]. Available: https://www.kdnuggets.com/2020/06/build-dog-breeds-classifier-aws-sagemaker.html. [Accessed: 01-May-2021].