**Friday March 08- with Dr.Diggans (mentor) & Professor Chen (instructor)**

- Last week: Evenly subsample from each category and extract intermediate feature spaces and singular vectors.

Variation to the original approaches (Dr.Diggans):

- Train the NN just on the subset (like 5000) to get some feature space representation. (might not be the best, but still serves as our embedding function), then feed forward the whole training set and look at the SVD of the whole training dataset. Apply the same procedure to subsamples.
- Using higher dimensional feature space (not just 64, but higher like 256) gives us more details on the features. This feature space encodes some data structure in the dataset.
- It might be better to break down by the class. For example, during the subsample phase, we take all the car images from the training set, do SVD on that and get all the feature spaces on car. Then take subsample from that class and compare the SVD of the subsample to the SVD of all the images in that class. Therefore, for each class, we end up with good subsamples, then use those good subsamples from each class by combining them together to form the new training subset.

But for now, stick with the original approach. Train CNN with whole training set of 45000, this serves as the embedding function. Use it to extract the feature space from whole dataset and multiple subsamples. If the feature spaces of certain subsamples are similar to the whole dataset one, then they are expected to perform well in training ML models like the original whole dataset.

Next week:

- We need 200 subsamples and each of them contains 5000 images (for now).
- Apply SVD and BSIE on them.
- Once we find a good subset (a good representation of the original dataset in terms of geometric distribution), then train a separate NN on that subset and compare the performance of that newly trained NN with the NN trained with the whole dataset on the test data (that neither NN has seen it yet).