

Thursday February 29 - with Dr.Diggans (mentor)

- Last week: using sample CNN to extract intermediate feature layer before the final classification layer for each image, and therefore form a matrix for the whole Cifar-10 training dataset. Its subsamples follow the same procedure as well.

Questions on CNN structure (Dr.Diggans):

- Why does CNN have 3 fully connected layers in the end? Goes from 1024 to 512 to 64 etc. Why cannot it just go from 1024 to 64?

Concerns (Dr.Diggans):

- The goal of the project is to use less data to train ML models, but to get the feature space of subsamples, we have to train on the whole dataset anyway (to attain a CNN model serving as our embedding function). Only with this embedding function, we can extract the intermediate feature space of the whole dataset and its subsamples. For example, to find out whether a subsample of 10000 is a good representation of the whole dataset, our approach is train the 45000 (the whole training dataset) and get the feature space of 10000, but the problem is that we already have trained on the whole dataset.
- Justification: train the whole dataset, and if that model is gone for some reason, and is there a subsample (given this must be a good representation of the whole dataset), then can regain the model quickly by training this subsample.

Other comments:

- Singular vectors extracted from matrices: one direction with the largest variance of the dataset (very large singular value), then exponentially drop off for the remaining ones.

Next Week:

- Extract the feature space of randomly selected subsamples using the trained CNN (Randomly sample from each class equally) and extract singular vectors of the subsample.
- Fix BSIE code.
- Pretrained CNN model to extract feature layers. (Using the existing CNN architecture)