Friday March 22- with Dr.Diggans (mentor)

Last week:

Even if increasing the subsample size (to 20000) to train the models, the test accuracy of models still lacks a correlation with the entropy values of those subsamples.

Dr.Diggans:

Go through each class (each class has 5000 training images) and subsample 1/10 of images from each class. Compute BSIE for each entire class and BSIE of the subsample from each class. If the entropy value of one subsample is close to the entropy value of the corresponding class, then it means that subsample is representative of that class. Then the ML models trained based on the subsample are expected to have better classification performance on the test set of that class. We will repeat this process for the whole 10 classes in Cifar-10 dataset.

To measure closeness of subsample entropy to the class entropy, we can compute the absolute value of BSIE of one class – BSIE of corresponding subsample from that class / BSIE of this class) and we repeat this process for each of the 10 classes.

When we take 100 subsamples from class1, some of them will be good and have smaller relative entropy errors and some of them are bad and have larger relative errors. In each round, we take a random sample from class1, a random sample from class2, and etc. up to class10. It could be the case that the subsample from class 1 is good, but bad subsamples for the rest of them. We could have 10 subsamples that have low relative errors for all 10 classes, but unlikely because this means all the classes are represented well. We then combine the subsamples to train CNN, and we test the model class by class. How good is this model is classifying cars, and compare with the relative entropy error for cars, then next object in Cifar-10 until we test our model in all 10 classes separately.

These subsamples we collect from each class are combined to train the model, and we are interested in what is its performance (test accuracy) on class1 compared to the relative entropy error of class1. (That is, just test it on the test data for class1)

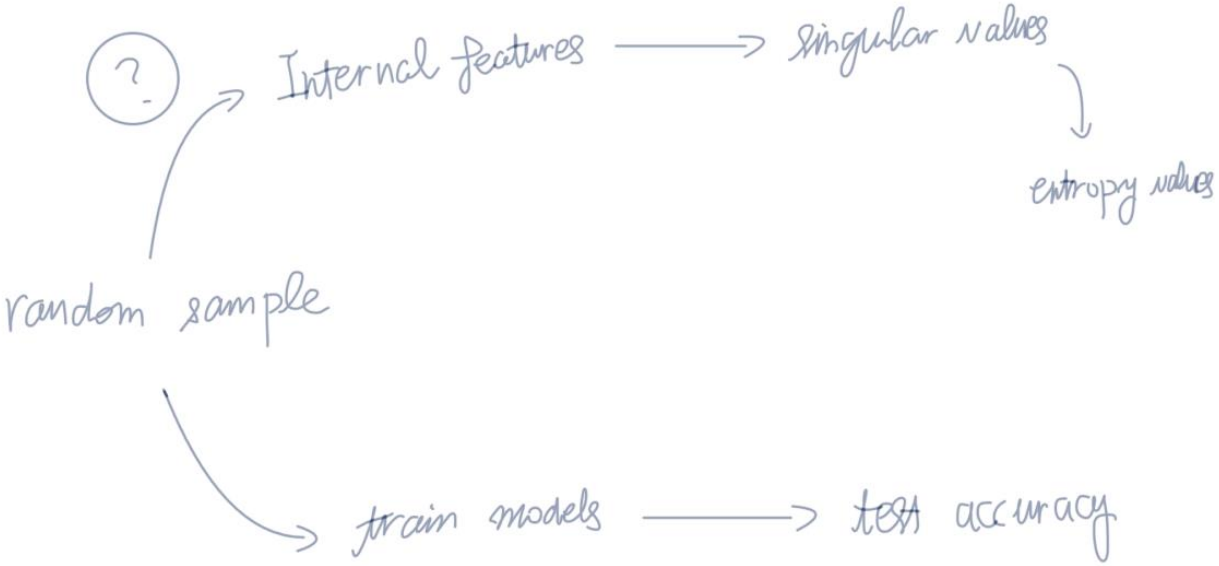
Don't look at the test accuracy for the whole test data

10 numbers of test accuracy for each class.

10 entropy relative errors for those 10 classes.

Other ideas:

- **Switch column orders won't affect singular values, which means entropy value won't change.**
- **Use pretrained to extract features and singular values and entropy values.**
- **Use the raw images e.g., 28*28*3 and convert it to a long vector (number of rows) * 50000 number of columns (use this as my feature space without letting it go through any models. Since CNN is a black box function, is it really extracting the features we want?) --- So try raw RGB values**
- **CNN in general condenses information**



Why is the original approach not working?

- **The problem is more likely in the upper route.**
- **Truncate the first few largest singular values (from middle singular vector after SVD) with respect to singular vectors of both the subsamples and whole training set and compute their entropy value as the rest of them are close to 0 and noisy.**

Next week:

Use the class-by-class approach to figure out if there is some correlation between the test set accuracy and relative entropy error of each class.