**ENM531 FInal Project Proposal**

Group Members: Zesheng Liu, Qianxun Xu

**An overview of the proposed problem and why it is important.**

In problem 2 from homework 4, we only get an accuracy of about 64%, which is quite low. Inspired by that problem, the goal of this project is to build and compare neural network models for image classification tasks that can achieve above 85%-90% accuracy, and learn the process of how to change our model, and increase the accuracy step by step.

The dataset we will use is the CIFAR-10 dataset, which consists of 60,000 32x32 color images of 10 classes, 6,000 images per class. We believe this task is both closely relevant to the topics covered in class, such as deep learning, neural networks and CNN, and relevant to the applications in computer vision, such as medical image analysis and facial recognition. As long as we try to get a better accuracy, we can compare the different architectures, training time, and performance of each model, and to understand the reasons why one model may outperform another. This analysis will help us understand the advantages of different architectures, thus we can build better models in the future.

**What are the challenges to be addressed?**

The first challenge is that the training dataset we selected is not particularly large. And to train on a large-scale dataset will require much more computational power to do so. Secondly, we are not familiar with the architectures and programs of other models we plan to implement yet, so it might be challenging for us to identify the key factors that induce the differences in the outcomes.

**What is the current state-of-the-art and which are its potential limitations.**

From this website (https://paperswithcode.com/sota/image-classification-on-cifar-10), we can see that the ViT-H/14 model, which uses Vision Transformer, achieves 99.5 accuracy on this dataset. ViT-H/14 applies a pure transformer directly to sequences of image patches and has shown impressive results on image classification tasks. Compared to state-of-the-art convolutional networks, ViT-H/14 requires less computational power and can arrive at a similar level of results.

When applying the ViT-H/14 model to CIFAR-10 dataset, there are a couple limitations: (1) To achieve 99.5% accuracy, ViT-H/14 was pre-trained on large-scale datasets. However, since our dataset is not particularly large, we may expect lower accuracy; (2) The model trained for one image processing task may not adapt well to other related tasks; (3) Visualization and interpretation of transformer models. (<https://arxiv.org/pdf/2101.01169.pdf>)

**What is the formal mathematical definition of the problem, and what is the technical approach you will employ.**

For the problem, It is an image classification task. We have several image data, which is in shape of 32\*32\*3, and we hope to make a prediction for them in 10 different classes. The basic mathematical definition for a simple CNN architecture iis that we will use convolution calculations to identify features from raw data and then pass into some activation functions and pooling layers, after several calculations we can flatten the 2D feature map into 1D and use a simple MLP network to do the classification.

We will keep the basic idea of CNN and try to do some data pre-processing or model modifications in order to get better accuracy at first, then we will try some other common image classification model and find out the model with better accuracy and less runtime.

**What is the potential impact upon the successful completion of the project?**

If we can successfully complete this project as our original plan, it will be a great experience for us to learn the difference between those common models for image classification. This project will also be extremely informative and foundational for us who plan to take courses focusing on computer vision. Furthermore, we can learn the process of starting with a baseline CNN model, then building steps and observing how it improves model accuracy.

**Basic Plan for what will be done in the project.**

Generally, if we want to increase the accuracy of doing classification for images compared to the model we are using now , there are two aspects.

One aspect is that we can **keep the network architecture** we are using now, but do some modification for that. First we can do data augmentation before we feed the training data into our model. We plan to try some of the most common method:

A.Mirroring/Rotation the raw data

B.Converting into grayscale

C.Random crop

Second, we may make some changes to our model but still keep the general architecture.

A.Use batch normalization

B.Dropout

C.Weight Decay and Regularization

D.Try to change the learning rate while training, as we may start with a higher learning rate and the after some iterations we can then change to use a lower one

The Other aspect is that we change to **some other architecture** to train the data. Some possible model that we will use are:

A. Other architecture of CNN, like VGG16 architecture or a little change from their architecture

B. ResNet(Both ResNet 50 & ResNet 18)

C. SImple MLP/MLP-mixer

D. Vision transformer

Models may change if we find that we really meet a big challenge on implementation, or for some models we may just cite other’s work and results.

At last, we will briefly compare between these model implementations on accuracy, runtime and number of parameters.