# CSED703B Vision and Language Assignment 1

Gyunam Park
20172050
IME, POSTECH
gnpark@postech.ac.kr

## 1. Introduction

MNIST dataset has been utilized for developing learning or pattern recognition algorithms due to its completeness coming from proper preprocessing and formatting. It is composed of 60,000 training instances and 10,000 test instances, each of which is labeled. In this report, I applied Convolution Neural Network(CNN), which is widely used for visual recognition tasks in recent time, to classify images into correct label. I've developed four different CNN variants and evaluated their performance.

## 2. Backgrounds

CNN is a sequence of layers, where differentiable function transforms one activation to another. Type of layers are as follows: Convolution Layer, Pooling Layer, Fully-connected Layer. These layers build a CNN architecture. Let's say input is 32x32x4 (i.e. width x height x channel) pixel values. Convolution Layer works as sliding window of the input; the layer calculates dot product of every region it slides and the weights it contains. If the layer is composed of 8 filters and 4x4-sized window, the resulting activation volume will be 14x14x8 with stride of 2. Pooling layer downsamples the dimension of activation value, in this case resulting in 7x7x8 (e.g. 2x2 pooling layer). Fully-connected layer will calculate probability-like scores of each label class. It is an ordinary Neural Networks where each neuron in a input layer is connected with all neurons in a output layer.

## 3. Experiments

I used CNN in order to classify hand-written digits correctly. In order to find a way to improve the model, I developed four different models where I varied the size of neural network, regularization method, and optimizer. In order to evaluate the effectiveness of each approach, I compared the error rate between baseline model and the others. To this end, I fixed number of epochs, performance metric, loss function, and activation as follows: 10, 'accuracy', 'logarithm loss', 'relu'

Baseline model is composed of a convolution layer and a pooling layer, each of which is composed of 16 5x5 filters and 2x2 filter, respectively. Then, dropout is applied for a regularization purpose. The fully-connected layer has 64 neurons and output layer consists of softmax function.

Model 2 is different from baseline model in its size of network. It contains two layers of convolution, pooling, and fully-connected layer, which indicates it is deeper than the baseline. It is also wider since convolution layers have 124 and 64 filters, respectively and the fully-connected layers are composed of 128 and 64 neurons.

model 3 utilize 'RMSprop' as optimizer rather than 'Adam'. Another setting is all the same with the baseline model.

model 4 applies 'Batch Normalization' for the regularization purpose. Batch Normalization facilitates the learning by normalizing inputs of each layer.

Table 1. Evaluation Result

| Model | # Conv | Optimizer | Regularization | Error |
|---|---|---|---|---|
| Baseline | 1 | Adam | Dropout | 0.941% |
| Model 1 | 2 | Adam | Dropout | 0.804% |
| Model 2 | 1 | RMSprop | Dropout | 1.043% |
| Model 3 | 1 | Adam | BatchNorm | 1.514% |

The results are depicted in Table 1. Model 2, which has deeper and wider layers, perform best among others. Model 3, which applies Batch Normalization, performs worst with error rate around 1.5%.

## 4. Conclusion

In this article, I developed four different CNNs to classify hand-written digits. As presented in the Experiment, the accuracy of model increased as the model expanded. However, in this experiment, I have not identified the best combinations of various techniques and varied the hyper-parameters such as number of epochs. These should be dealt with in the future work.