University of Nebraska Lincoln

The Art & Science of Training Deep Neural Networks
CSCE 496/896

Programming Assignment 1

Spring 2021

Deep Study of Fully-Connected Feedforward Deep Neural Networks

---

CSCE 496: 110 points
CSCE 896: 155 points

Last Name 1: Gao

First Name 1: Tengjun(David)

NUID 1: 62915237

Last Name 2: Choi

First Name 2: Sung Woo

NUID 2: 79920240

Obtained Score:

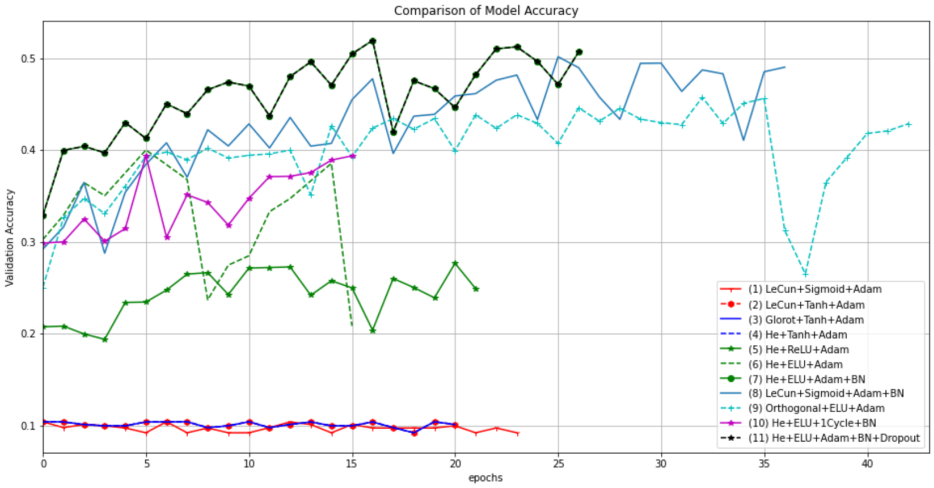
496 & 896 PA1 Report

Gao Tengjun(David)
david.gao313@huskers.unl.edu

Sung Woo Choi
schoi9@huskers.unl.edu

1/9/2021

1 A single graph showing learning curves of all experiments: Epochs vs. validation accuracy



2 A single table with the following columns showing comparison of all experiments.Experiment No.-Number of Epochs-Training Time(wall time)-Test Accuracy

Experiment No.	Name	No. of Epochs	Training Time (wall time)	Test Accuracy
1	LeCun + Sigmoid + Adam	24	3 min 56 s	0.1000
2	Lecun + Tanh + Adam	21	3 min 30 s	0.1000
3	Glorot + Tanh + Adam	21	3 min 32 s	0.1000
4	He + Tanh + Adam	21	3 min 29 s	0.1000
5	He + ReLU + Adam	22	3 min 38 s	0.2715
6	He + ELU + Adam	16	2 min 50 s	0.4094
7	He + ELU + Adam + BN	27	8 min 16 s	0.5121
8	LeCun + Sigmoid + Adam + BN	37	10 min 47 s	0.4897
9	Orthogonal + ELU + Adam	43	7 min 40 s	0.4699
10	He + ELU + 1cycle + BN	16	3min 59s s	0.4113
11	He + ELU + Adam + BN + Dropout	27	8 min 16 s	0.5121

3 Answer the following questions

1. Which experiment performed the best(test accuracy and training time)? Why?

   For test accuracy, experiment 7 (He + ELU + Adam + BN) and experiment 11 (He + ELU + Adam + BN + Dropout) performed the best. They both resulted with 0.5121 test accuracy because they have He initializer, ELU activation function, and Adam optimizer. Although experiment 11 has Dropout that deactivates neurons in the layers randomly based on the drop out rate in order to overcome the data over-fitting, Dropout did not affect at all in our experiment 11. Experiments 7 and 11 performed the best in test accuracy because both experiments use He weight initializer and ELU activation which are great combination that resolves the vanishing and exploding gradient problem. Furthermore, they have batch normalization which is an effective technique to avoid vanishing gradient problem by ensuring that the weighted signals have small enough variance not to cause saturation.

   For training time, experiment 6 (He + ELU + Adam) performed the fastest. The training time of this experiment was 2min 50s. It resulted in 16 epochs which is the smallest number of epochs among all experiments. Since this experiment does not have batch normalization, computation time of each epoch resulted between 7 and 9 ms/step. Thus, it performed the fastest in training time. However, it is one of the experiments with the worst test accuracy.

2. Which experiment performed the worst(test accuracy and training time)? Why?

For test accuracy, experiment 1 (LeCun + Sigmoid + Adam), experiment 2 (LeCun + Tanh + Adam), experiment 3 (Glorot + Tanh + Adam) and experiment 4 (He + Tanh + Adam) performed the worst. They resulted with 0.1 test accuracy. Except experiment 8 (LeCun + Sigmoid + Adam + BN), these are the methods that have logistic sigmoid/tanh activation. Experiment 8 still resulted much better with sigmoid activation than these four methods because it had batch normalization. Since these four methods have logistic sigmoid/tanh activation, they resulted very poorly in test accuracy compared to other experiments.

The range of tanh will always lie between [-1,1], sigmoid is compressing outputs to lie between[0,1]. Their ranges are different only by 1 and their shapes in the graph are very similar, which mean they will behave almost the same when encountering the large data. The reason why logistic sigmoid/tanh activation results in low test accuracy is because of 'vanishing gradient problem' in both activation. The output will not get updated as much in the end which will hardly match the expected output.

Overall, the experiment 1 performed the worst in test accuracy and training time. Experiment 1 has 2 or 3 more epochs compared other experiments with the worst test accuracy which led longer training time.

For training time, experiment 8 (LeCun + Sigmoid + Adam + BN) resulted with the worst/longest training time. It took 10min 47s of training time. Since experiment 8 has batch normalization, computation time of each epoch resulted in 13 ms/step which is the longest computation time per epoch among all experiments. Also, the experiment 8 resulted with 37 epochs, which is the second greatest number of epochs resulted among all experiments. Thus, the experiment 8 resulted with the longest training time.

3. Compare experiment 6 with experiment 7? Analyze the results and report your observations. Analyze the computational overhead of BN (per epoch time increase).

   Experiment 6 (He + ELU + Adam) resulted in 16 epochs with 2min 50s of training time and 0.4094 test accuracy. Experiment 7 (He + ELU + Adam + BN) resulted in 27 epochs with 8min 16s of training time and 0.5121 test accuracy. The batch normalization caused longer computation time of each epoch. It took an average of 3 ms/step more computation time per epoch. For experiment 6, each epoch took between 7 ms/step and 9 ms/step. For experiment 7, each epoch took 13 ms/step. Since the experiment 7 has longer computation time per epoch and greater number

of epochs compared to the experiment 6, the training time resulted longer for experiment 7.

Since experiment 7 with BN can handle better with vanishing gradient problem, early stopping happened 12 epochs later for experiment 7 compared to experiment 6. Thus, the model of experiment 7 is trained better and resulted with higher test accuracy compared to experiment 6.

4. Compare experiment 8 with experiment 1? Analyze the results and report your observations.

   Experiment 8 (LeCun + Sigmoid + Adam + BN) resulted in 10min 47 with 37 epochs and the test accuracy of 0.4897. Experiment 1 (LeCun + Sigmoid + Adam) resulted in 3min 56s with 24 epochs and the test accuracy of 0.1000. Experiment 8 is the experiment that resulted with longest training time in this assignment. Since experiment 8 has batch normalization, the time computation of each epoch of experiment 8 is longer than that of experiment 1: each epoch of experiment 1 is computed with 7 ms/step while each epoch of experiment 8 is computed between 12 and 13 ms/step.

   Regarding test accuracy, experiment 1 performed very poorly compared to experiment 8. Batch normalization normalizes (unit variance) the distribution of input in each layer, thereby avoids creating vanishing gradient. Since experiment 8 with BN can handle better with vanishing gradient, the early stopping happened 13 epochs later for experiment 8 compared to experiment 1. Thus, the model of experiment 8 is trained better and resulted with higher test accuracy compared to experiment 1.

5. Compare experiment 9 with experiment 6? Analyze the results and report your observations.

   Experiment 9 (Orthogonal + ELU + Adam) resulted in 7min 40s with 43 epochs and the test accuracy of 0.4699 while experiment 6 (He + ELU + Adam) resulted in 2min 50s with 16 epochs and the test accuracy of 0.4094. Although computation time of each epoch resulted between 7 and 8 ms/step for both experiments, experiment 9 resulted with greater number of epochs which lead to longer training time. Experiment 9 has greater number of epochs because orthogonal weight initalizer is better at handling vanishing and exploding gradients than He weight initialzer. Thus, early stopping happened in 16 epochs for experiment 6 while it happened in 43 epochs for experiment 9. As a result, the model of experiment 9 trained better than experiment 6 and resulted with higher test accuracy than experiment 6 with He weight initializer.

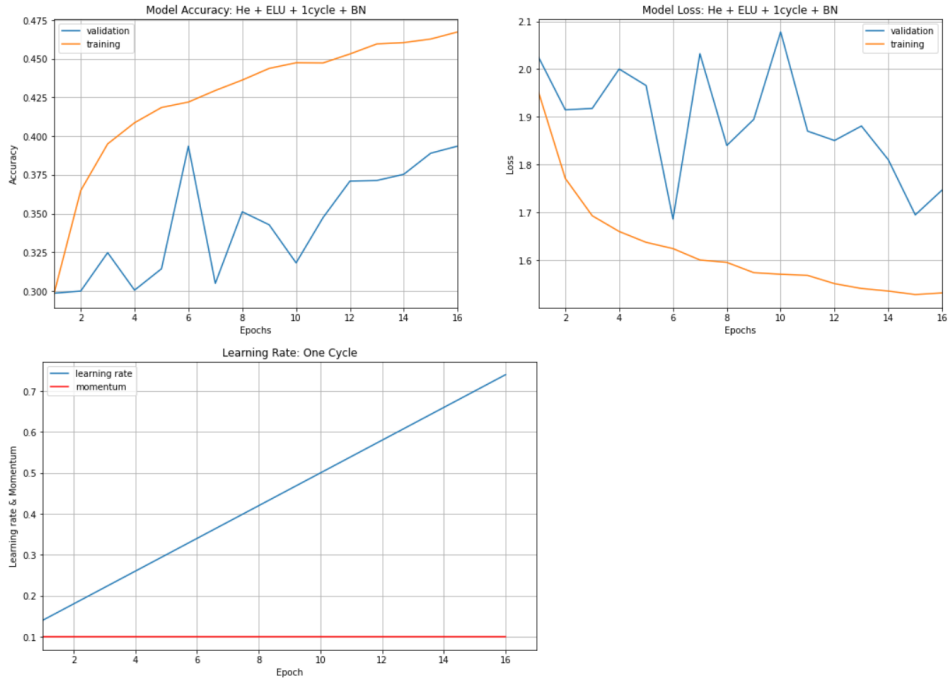
6. Compare experiment 10 with experiment 7? analyze the results and report your observations. How does 1cycle learning rate schedule change the training time?

   Experiment 10 (He + ELU + 1cycle + BN) resulted in 16 epochs with 3min 59s and 0.3988 training time. Experiment 7 (He + ELU + Adam + BN) resulted in 27 epochs with 8 min 16s and 0.5121 training time. The difference between experience 10 and 7 is that experiment 10 is experimented with 1 cycle schedule and SGD optimizer while experiment 7 is experiment with Adam optimizer.

   For finding the maximum learning rate for 1 cycle schedule in experiment 10, we used SGD optimizer with 0.1 learning rate and 0.1 momentum. Based on loss vs learning rate graph, we choice 1.0 as our maximum learning rate. However, compared with Smith and Toping 2017 paper, our maximum learning rate are very large that will cause over-fitting. Thus, we have not seen one cycle of up and down in learning rate momentum vs epoch graph due to early stopping.

```
Epochs:  16
He + ELU + 1cycle + BN - Train Accuracy:  0.4112222194671631
He + ELU + 1cycle + BN - Test Accuracy:  0.39879998564720154

He + ELU + 1cycle + BN - Train Loss:  1.641879916191101
He + ELU + 1cycle + BN - Test Loss:  1.6794155836105347
```



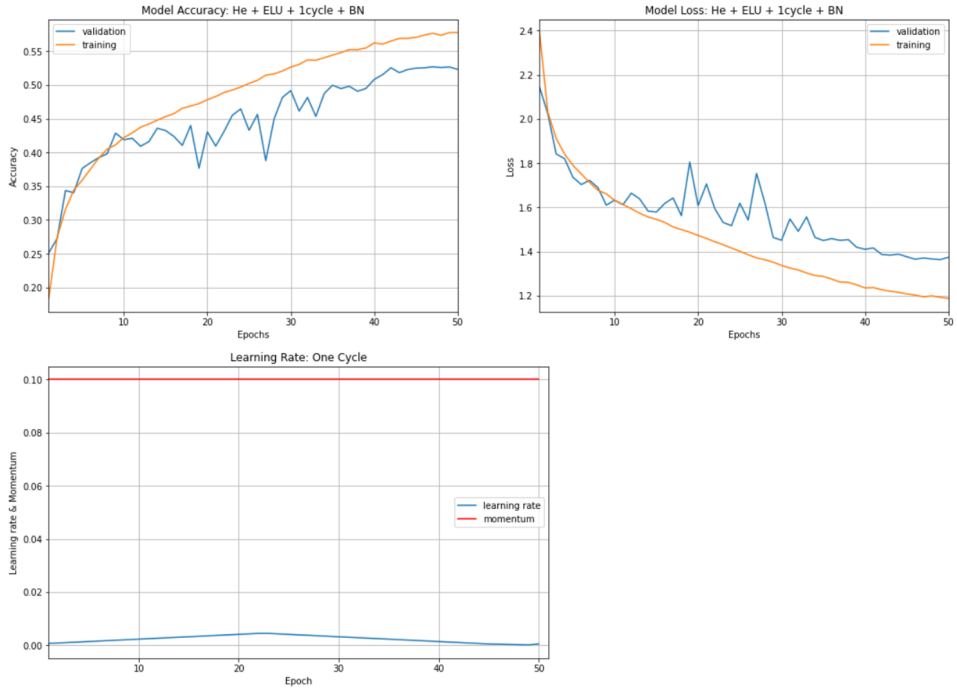
Experiment 10. (He + ELU + 1cycle + BN) with 1.0 max learning rate

Based on Smith and Toping in their 2017 paper, we found that their maximum learning rate seems 0.0045 based on their learning rate vs epoch graph. We have tried training our experiment with maximum learning rate of 0.0045 and got a result as below.

```
CPU times: user 18min 6s, sys: 39.2 s, total: 18min 46s
Wall time: 12min 37s

Epochs:  50
He + ELU + 1cycle + BN - Train Accuracy:  0.6592000126838684
He + ELU + 1cycle + BN - Test Accuracy:  0.522599995136261

He + ELU + 1cycle + BN - Train Loss:  0.9738677740097046
He + ELU + 1cycle + BN - Test Loss:  1.3829996585845947
```



Experiment 10. (He + ELU + 1cycle + BN) wih 0.0045 max learning rate

With maximum learning rate of 0.0045, we were able to see one cycle of up and down in learning rate momentum vs epoch graph. The test accuracy resulted with 0.5226 and training time ended in 12min 37s.

Since we were not able to find the optimal maximum learning rate based on our method, over fitting happened and thus early stopping executed. As a result, experiment 10 resulted in 16 epochs with 3min 59s while experiment 7 resulted in 27 epochs with 8min 16s. Based on our experiment, 1cycle learning rate schedule changed the training time of the experiment 10 about 10 percent faster than the training time of experiment 7.

7. Compare experiment 11 with experiment7? Analyze the results and report your observations.

Interestingly both experiments 11 (He + ELU + Adam + BN + Dropout) and 7 (He + ELU + Adam + BN) resulted with the same number of

epochs, training time, test accuracy, and train accuracy. Both experiments resulted in 27 epochs with 8min 16s of training time, 0.5121 test accuracy, and 0.6102 train accuracy. The advantage of using dropout is that at each iteration it temporarily drops neurons based on the dropout rate which prevents all neurons in a layer from synchronously optimizing their weights. In experiment 11, we experimented with 0.1 dropout rate. Based on the results, dropout did not affect anything on experiment 11. This could mean that dropout dropped the neurons with unrelated features that do not update the weights and bias. However, we are not sure why dropout does not affect in experiment 11.