**Name(s): Seung Wan Yoo**

**NetID(s): wanyoo2**

**Team name on Kaggle leaderboard: Wan Yoo**

**For each of the sections below, your reported test accuracy should approximately match the accuracy reported on Kaggle**.

*Briefly describe the hyperparameter tuning strategies you used in this assignment. Then record your optimal hyperparameters and test/val performance for the four different network types.*

NOTE\*\*: I forgot to keep precise track of the various hyperparameters I have altered and modified that got me the result of my Kaggle score, thus I readjusted and wrote the report.

**Two-layer Network Trained with SGD**

*Best hyperparameters (if you changed any of the other default hyperparameters like initialization method, etc. please note that as well):*

* *The two-layered SGD network’s increased significantly when I increased hidden size to 120 from the default value. Also, changes to learning rates, regularization coefficient caused major difference between results but the strategies of learning rate – decay was the one I spent the most time on. I tried decay by accuracy, where the decrease of validation accuracy compared to previous epoch which did not turn out successfully and constant decrease per 5 epoch, which also did not turn out well. I decided that percentage decrease every epoch yielded the best results. I wanted to see how the learning rate decayed thus I created a list to keep track of the learning rates. The number of iteration is reduced to 40 as it most often plateaus from near 35+~*

|  |  |
| --- | --- |
| Batch size: | 200 |
| Learning rate: | 0.05 |
| Hidden layer size: | 120 |
| Regularization coefficient: | 0.0004 |
| Learning rate decay rate: | 0.85 |
| Epoch: | 40 |

*Record the results for your best hyperparameter setting below:*

|  |  |
| --- | --- |
| Validation accuracy: | 0.523 |
| Test accuracy: | 0.5247 |

**Three-layer Network Trained with SGD**

*Best hyperparameters (if you changed any of the other default hyperparameters like initialization method, etc. please note that as well):*

* *The strategy was essentially the same for the two-layer SGD network. The hyperparameters that yielded the best results had lower regularization coefficients because they were fixed since the validation classification was decreasing. The number of iteration is reduced to 40 as it most often plateaus from near 35+~.*

|  |  |
| --- | --- |
| Batch size: | 200 |
| Learning rate: | 0.05 |
| Hidden layer size: | 120 |
| Regularization coefficient: | 0.0004 |
| Learning rate decay rate: | 0.85 |
| Epoch: | 40 |

*Record the results for your best hyperparameter setting below:*

|  |  |
| --- | --- |
| Validation accuracy: | 0.535 |
| Test accuracy: | 0.5288 |

**Two-layer Network Trained with Adam**

*Best hyperparameters (if you changed any of the other default hyperparameters like initialization method, etc. please note that as well):*

* *Training Network with Adam was different story than that of SGD update network. The update rule was different(exponential) and there were two additional hyperparameters to care about. The learning rate had to be lower due to possible overflow or infinite number.*

*The Adam update, although very quick to converge at such high rates and accuracy, It is not exactly the most stable model. The number of iteration is reduced to 40 as it most often plateaus from near 35+~.*

|  |  |
| --- | --- |
| Batch size: | 200 |
| Learning rate: | 0.003 |
| Hidden layer size: | 120 |
| Regularization coefficient: | 0.0001 |
| β1 | 0.6 |
| β2 | 0.69 |
| Learning rate decay rate: | 0.85 |
| Epoch: | 40 |

*Record the results for your best hyperparameter setting below:*

|  |  |
| --- | --- |
| Validation accuracy: | 0.529 |
| Test accuracy: | 0.5058 |

**Three-layer Network Trained with Adam**

*Best hyperparameters (if you changed any of the other default hyperparameters like initialization method, etc. please note that as well):*

* *The notes are similar to those of the two-layered Adam network, although I have made changes to the regularization coefficient and the learning rate decreased significantly. This allowed the validation accuracy to get closer to the test accuracy. The number of iteration is reduced to 40 as it most often plateaus from near 35+~.*

|  |  |
| --- | --- |
| Batch size: | 200 |
| Learning rate: | 0.0003 |
| Hidden layer size: | 120 |
| Regularization coefficient: | 0.00001 |
| β1 | 0.7 |
| β2 | 0.79 |
| Learning rate decay rate: | 0.85 |
| Epoch: | 40 |

*Record the results for your best hyperparameter setting below:*

|  |  |
| --- | --- |
| Validation accuracy: | 0.526 |
| Test accuracy: | 0.5322 |

**Comparison of SGD and Adam**

*Attach two plots, one of the training loss for each epoch and one of the validation accuracy for each epoch. Both plots should have a line for SGD and Adam. Be sure to add a title, axis labels, and a legend.*

*Compare the performance of SGD and Adam on training times and convergence rates. Do you notice any difference? Note any other interesting behavior you observed as well.*

*The following comparison was done on 3-layered Adam and sgd. The hyperparameters were the same.*

*Chart

Description automatically generated with medium confidence*

The train loss for Adam was much steeper for the Adam and it much lower than SGD model.

*Chart, line chart

Description automatically generated*

The validation accuracy for SGD rises little and slowly compared to the Adam, where it rises very steeply in the beginning and plateaus eventually. The validation accuracy for Adam rises to over 50% where as SGD with the same parameters rise barely over 40%.