Name(s): Theodoros Mamalis
NetID(s): mamalis2
Team name on Kaggle leaderboard: PeacockNavigator

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.

The batch size and number-of-epochs hyperparameters were picked so that the runtime of the algorithms is reasonable (in general, increasing the total-epochs value by a considerable (runtime-wise) amount increased the performance only slightly but without overfitting, at least for the values tried in the experiments below). The hidden size was picked to be 120 as per the recommendation included in the provided python files (nearby values were tried as well but underperformed). The regularization coefficient was picked through trial and error, by increasing or decreasing the constant by increments of 0.1 or 0.01. The same is true for picking the learning rate.

The nonlinearity used for all models was Relu for all layers besides the last, for which softmax was used. Moreover, the type of regularization used was L2 for all models.

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 learning rate η used was: $\eta = $ *Learning rate* $\cdot\ 0.98^k$, where *k* is number of epochs passed. The learning decay 0.98 was picked after trial and error, starting with the given 0.95 and increasing it by 0.01. Reducing it was avoided to prevent the step size from being too small.

Batch size:	2000
Learning rate:	0.1
Hidden layer size:	120
Regularization coefficient:	0.2

Record the results for your best hyperparameter setting below:

Validation accuracy:	49.2
Test accuracy:	48.85

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 learning rate η used was: $\eta = \textit{Learning rate} \cdot 0.98^k$, where k is number of epochs passed. The learning decay 0.98 was picked after trial and error, starting with the given 0.95 and increasing it by 0.01. Reducing it was avoided to prevent the step size from being too small.

Batch size:	2000
Learning rate:	0.1
Hidden layer size:	120
Regularization coefficient:	0.2

Record the results for your best hyperparameter setting below:

Validation accuracy:	47.0
Test accuracy:	46.37

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):

The learning rate η used was: $\eta = 0.5 \cdot \textit{Learning rate} \cdot (1 + cos(t \cdot \pi / K))$ where t is number of iterations passed, and K is total number of epochs.

Batch size:	3000
Learning rate:	0.001
Hidden layer size:	120
Regularization coefficient:	0.01
β_1	0.9
β_2	0.999

Record the results for your best hyperparameter setting below:

Validation accuracy:	48.6
Test accuracy:	47.91

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 learning rate η used was: $\eta = 0.5 \cdot \textit{Learning rate} \cdot (1 + \cos(t \cdot \pi / K))$ where t is number of iterations passed, and K is total number of epochs.

Batch size:	3000
Learning rate:	0.001
Hidden layer size:	120
Regularization coefficient:	0.01
β_1	0.9
β_2	0.999

Record the results for your best hyperparameter setting below:

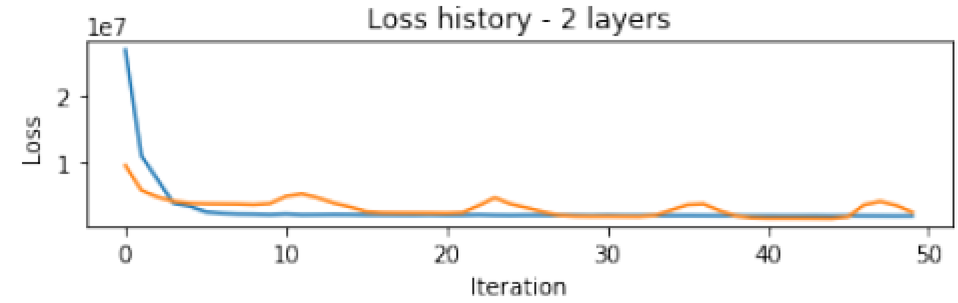
Validation accuracy:	53.7
Test accuracy:	54.82

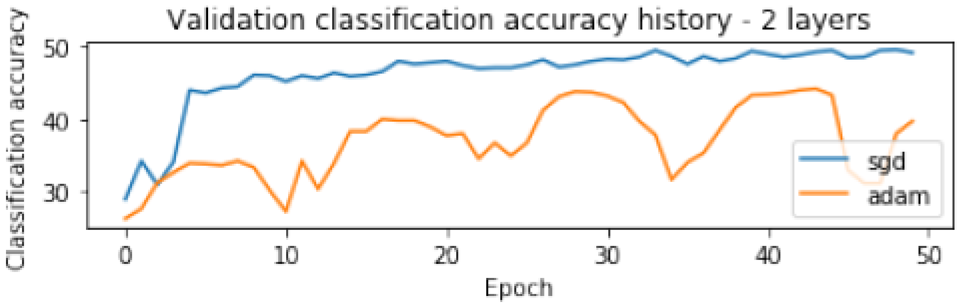
Comparison of SGD and Adam

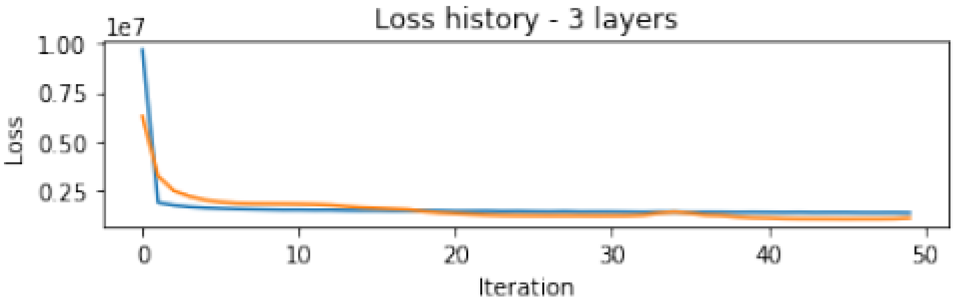
Attach two plots, one of the training loss for each epoch and one of the validation accuracy for each epoch. Both plots should have a line for SGD and Adam. Be sure to add a title, axis labels, and a legend.

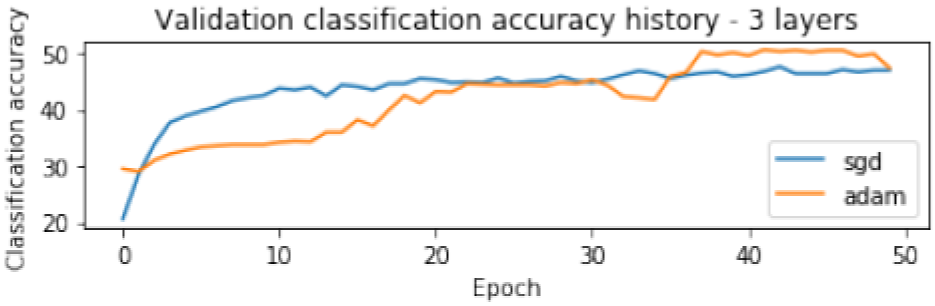
Compare the performance of SGD and Adam on training times and convergence rates. Do you notice any difference? Note any other interesting behavior you observed as well.

The plots are included below. It can be observed that SGD starts from a higher training loss for both the 2- and the 3-layer case. Moreover, SGD is faster to reach a higher validation accuracy. The repeated pattern in the ADAM case is because a sinusoidal learning rate was used instead of the exponentially decreasing learning rate in the SGD case. Finally, in the 3-layer case, ADAM is able to surpass SGD before 50 epochs, accuracy-wise.


Loss history - 2 layers
1e7
Loss
Iteration


Validation classification accuracy history - 2 layers
Classification accuracy
Epoch
sgd
adam


Loss history - 3 layers
1e7
Loss
Iteration


Validation classification accuracy history - 3 layers
Classification accuracy
Epoch
sgd
adam