Name(s): Marzuk Rashid
NetID(s): marzukr2
Team name on Kaggle leaderboard: Marzuk Rashid

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.

Two-layer Network Trained with SGD

For batch size, I tried 50, 100, and 200. I found 100 to work the best while not being too slow. For the learning rate, I tried 1e-3, 1e-2, and 1e-1. I found 1e-2 to be the best balance between too fast and too slow. For the hidden layer size, I tried 20, 70 and 120. I found a size of 120 to work the best. For the regularization constant, I tried 0.01, 0.05, and 0.1. I found 0.05 to work the best.

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

Batch size:	100
Learning rate:	1e-2
Hidden layer size:	120
Regularization coefficient:	0.05

Record the results for your best hyperparameter setting below:

Validation accuracy:	0.493
Test accuracy:	0.49770

Three-layer Network Trained with SGD

For batch size, I tried 50, 100, and 200. I found 50 to work the best. For the learning rate, I tried 1e-3, 1e-2, and 1e-1. I found 1e-2 to work the best. For the hidden layer size, I tried 20, 70 and 120. I found a size of 70 to work the best. For the regularization constant, I tried 0.01, 0.02, and 0.04. I found 0.04 to work the best.

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

Batch size:	50
Learning rate:	1e-2
Hidden layer size:	70
Regularization coefficient:	0.04

Record the results for your best hyperparameter setting below:

Validation accuracy:	0.489
Test accuracy:	0.49290

Two-layer Network Trained with Adam

For batch size, I tried 50, 100, and 200. I found 100 to work the best. For the learning rate, I tried 1e-4, 2e-4, 4e-4, 1e-3, 1e-2, and 2e-4. I found 2e-4 to work the best. For the hidden layer size, I tried 20, 70 and 120. I found a size of 120 to work the best. For the regularization constant, I tried 0.01, 0.02, and 0.04. I found 0.02 to work the best. For β_1 I tried 0.9, 0.99 and 0.999. I found 0.9 to work the best. For β_2 I tried 0.99, 0.999, 0.9999, 0.99999, and 0.999999. I found 0.99999 to work the best.

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

Batch size:	100
Learning rate:	2e-4
Hidden layer size:	120
Regularization coefficient:	0.02
β_1	0.9
β_2	0.99999

Record the results for your best hyperparameter setting below:

Validation accuracy:	0.531
Test accuracy:	0.52240

Three-layer Network Trained with Adam

For batch size, I tried 50, 100, and 200. I found 50 to work the best. For the learning rate, I tried 1e-4, 2e-4, 4e-4, 1e-3, 1e-2, and 2e-4. I found 2e-4 to work the best. For the hidden layer size, I tried 20, 70 and 120. I found a size of 70 to provide enough granularity without overfitting. For the regularization constant, I tried 0.01, 0.02, and 0.04. I found 0.02 to work the best. For β_1 I tried 0.9, 0.99 and 0.999. I found 0.9 to work the best. For β_2 I tried 0.99, 0.999, 0.9999, 0.99999, and 0.999999. I found 0.99999 to work the best.

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

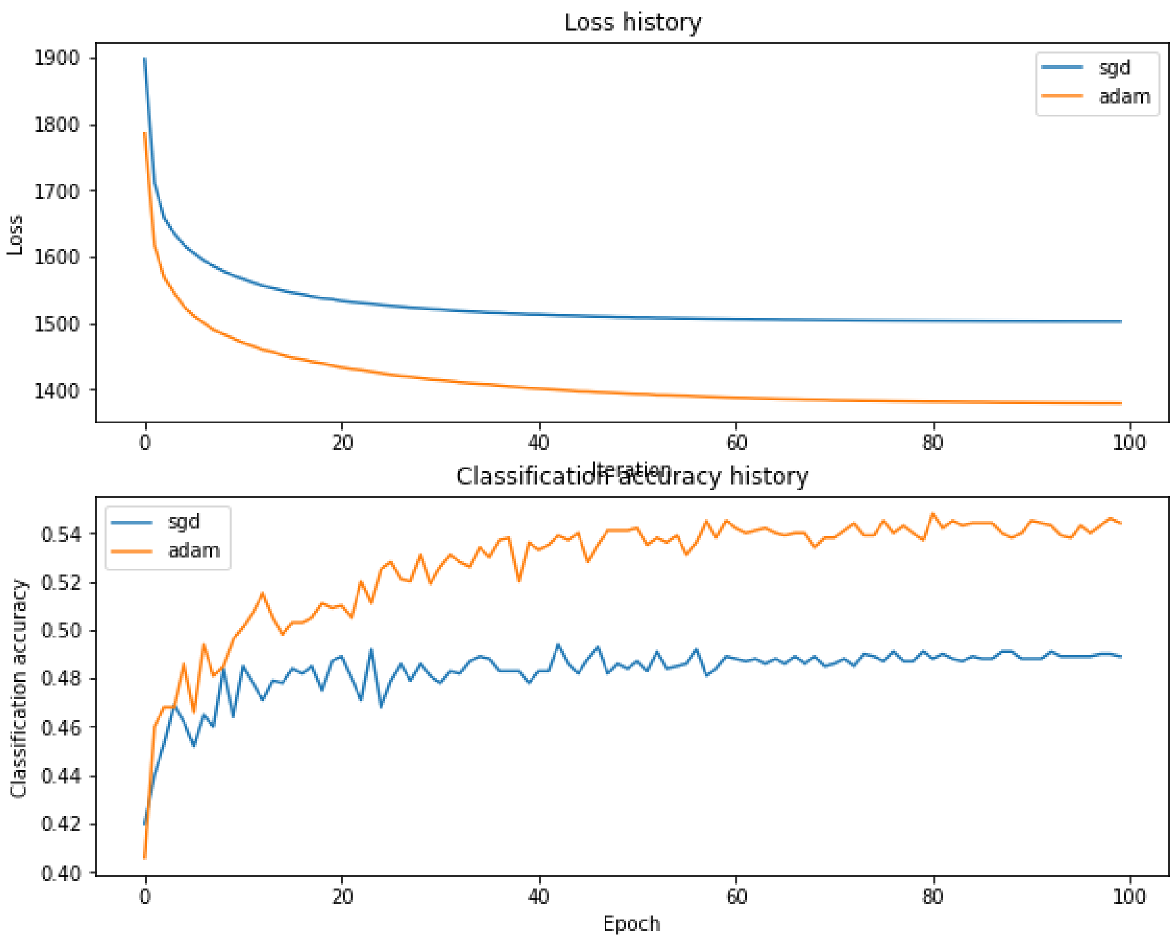
Batch size:	50
Learning rate:	2e-4
Hidden layer size:	70
Regularization coefficient:	0.02
β_1	0.9
β_2	0.99999

Record the results for your best hyperparameter setting below:

Validation accuracy:	0.544
Test accuracy:	0.52540

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.

Adam converges much faster than SGD and is able to converge to a better solution. This is shown both in the lower loss and higher validation accuracy. Apart from this difference, SGD seems to be similar to Adam based on these plots.