

[Clear selection](#)[Share](#)[Comment](#)[Star](#)

Preethi MM21B051 DA6401 - Assignment 1

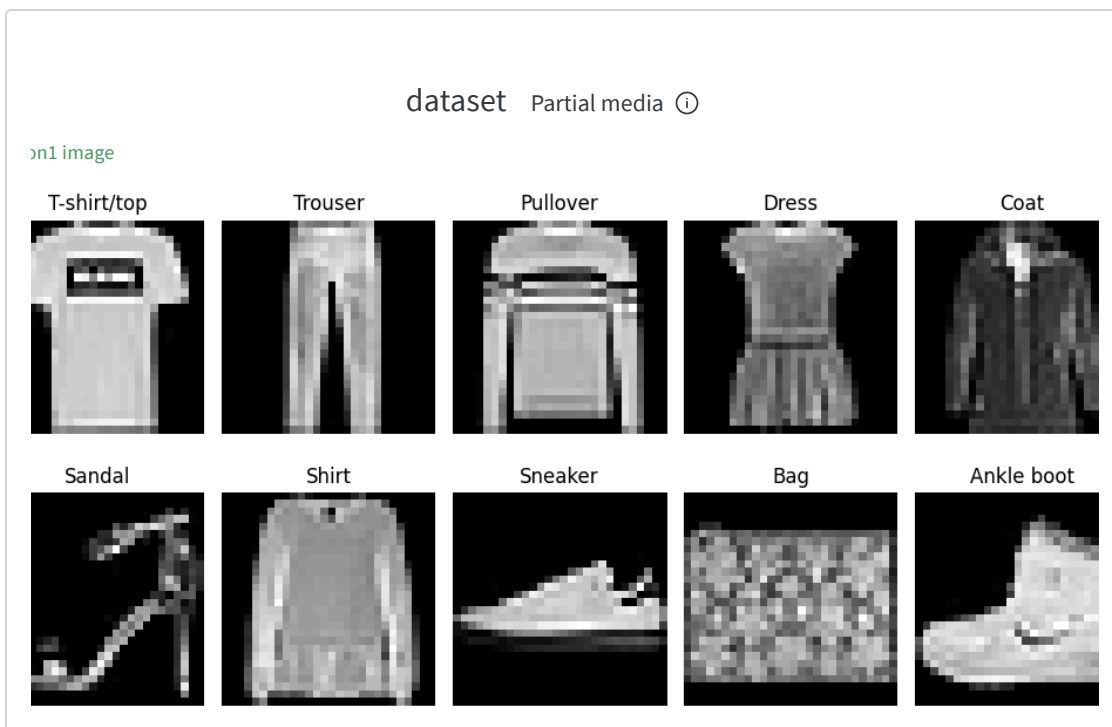
Link to the wandb report <https://wandb.ai/mm21b051-iitmaana/DeepLearning1/reports/Preethi-MM21B051-DA6401-Assignment-1--VmlldzoxMTgzODA3OA?accessToken=7wlbkqfgeb1v5oe0mfcxpvt8pryu6qw69xp5gp1nzjvxsxo...>

Preethi B mm21b051

Created on March 17 | Last edited on March 17

▼ Answer 1 (2 Marks)

Downloaded the fashion-MNIST dataset and plotted the first image of each class





Run set 2 448 ⋮

▼ Answer 2 (10 Marks)

Implemented a feedforward neural network which takes images from the fashion-mnist data as input and outputs a probability distribution over the 10 classes.

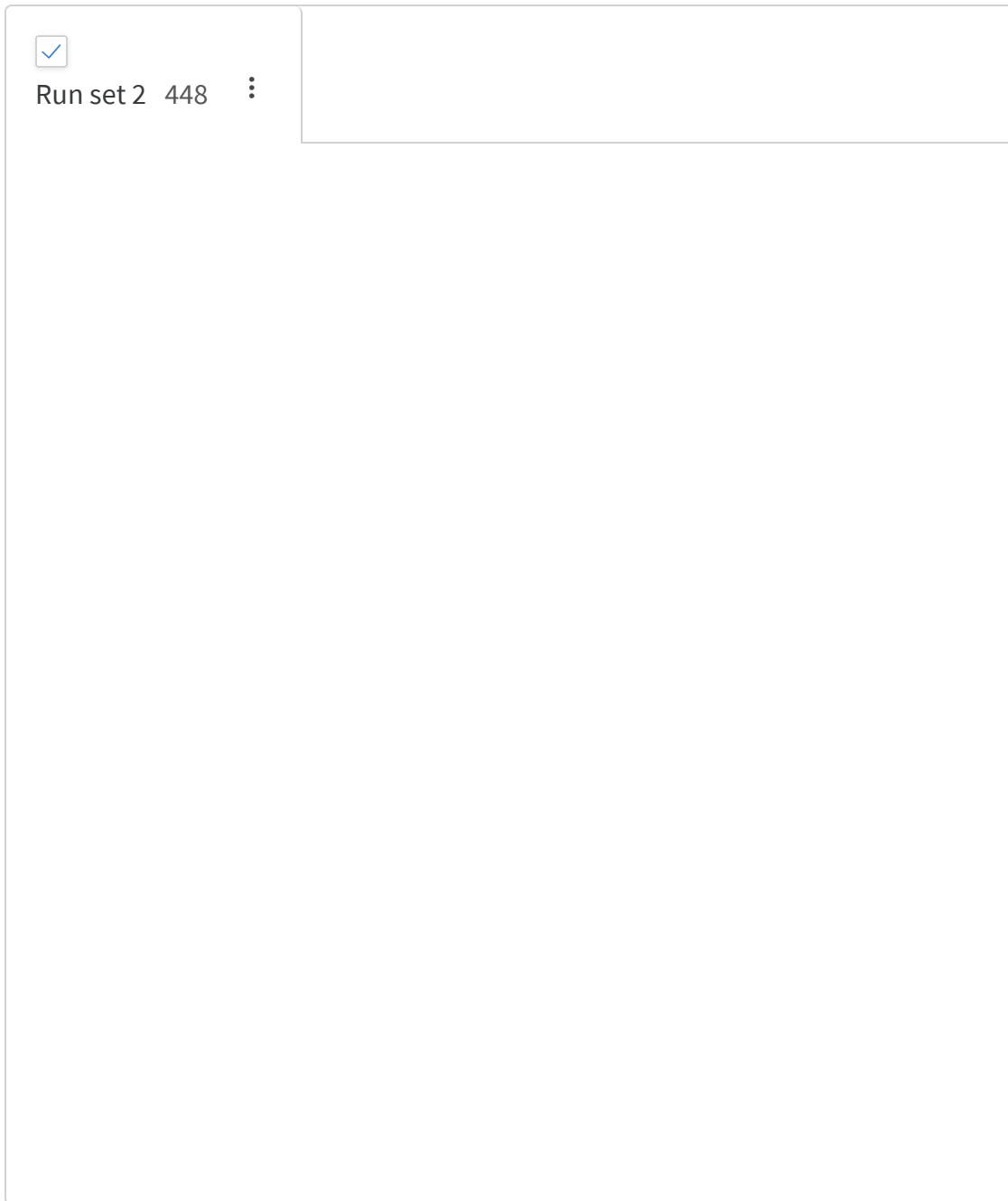
The number of hidden layers and the number of neurons in each hidden layer can be changed, along with other parameters of the model.

▼ Answer 3 (24 Marks)

Implemented the backpropagation algorithm with support for the following optimisation functions

- sgd
- momentum based gradient descent
- nesterov accelerated gradient descent
- rmsprop
- adam
- nadam

We see that rmsprop has the lowest loss of 0.13 followed by adam optimizer with a loss of 0.23



▼ Answer 4 (10 Marks)

Ran sweep with the following hyper parameters and logged them on wandb

- number of epochs: 5, 10
- number of hidden layers: 3, 4, 5
- size of every hidden layer: 32, 64, 128
- weight decay (L2 regularisation): 0, 0.0005, 0.5
- learning rate: 1e-3, 1 e-4
- optimizer: sgd, momentum, nesterov, rmsprop, adam, nadam
- batch size: 16, 32, 64
- weight initialisation: random, Xavier
- activation functions: sigmoid, tanh, ReLU



Run set 2 100 ⋮

▼ Answer 5 (5 marks)

Scatter plot of validation accuracy for all the models

Run set 2 100 ⋮

▼ Answer 6 (20 Marks)

Through the correlation plot we see that the optimizer used has high importance in the val accuracy, further adam optimizer has the best positive correlation indicating that it is the best optimizer to be used. This is further supported in the parallel co-ordinates plot, adam and rms-prop have the best results.

Poorly performing model configurations (below 65%):

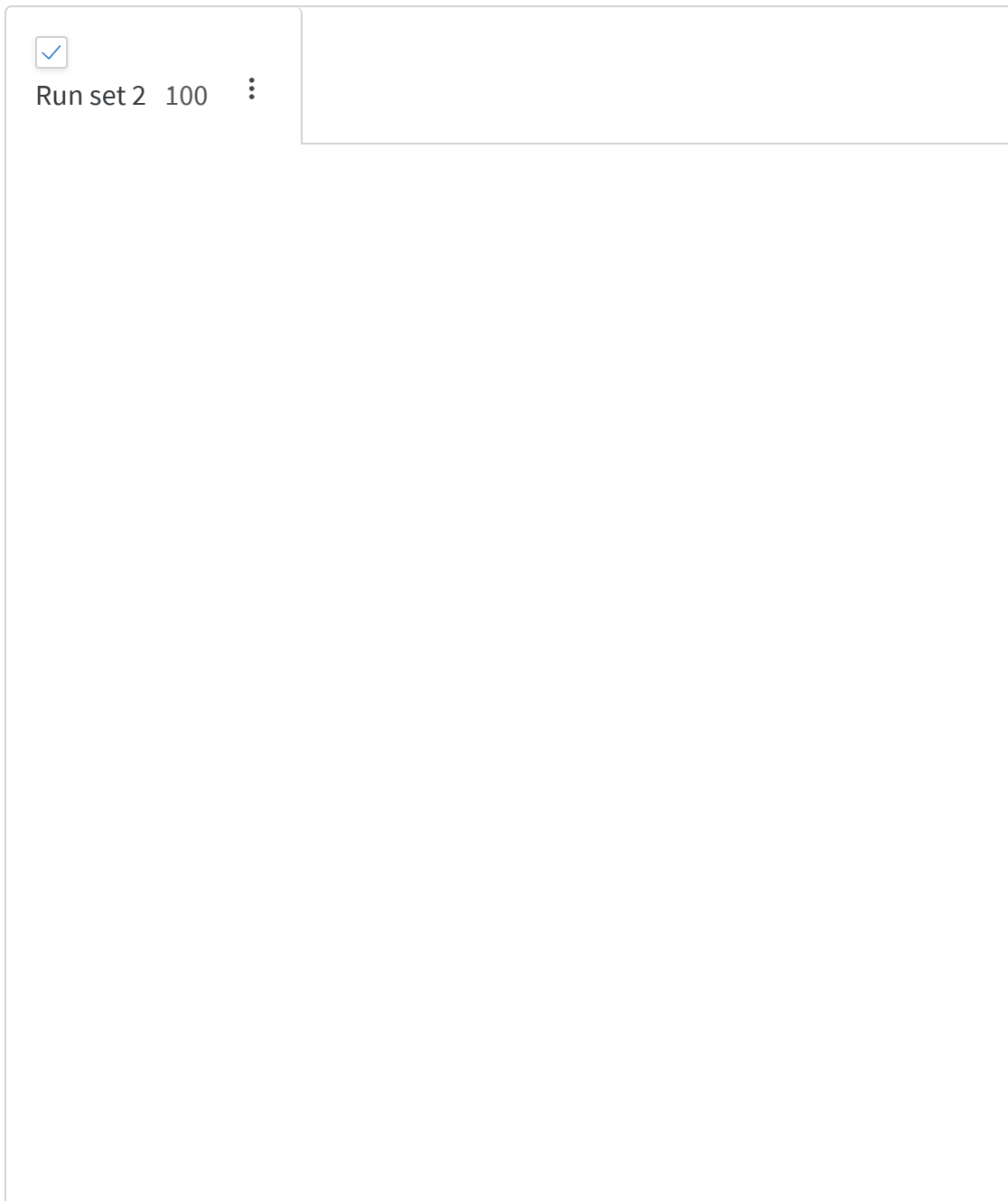
- most of them have sigmoid activation, few of other activation like tanh or relu
- batch size doesn't seem to have much importance on the val_accuracy compared to the other parameters, but lower batch sizes seem to have lower accuracy
- epochs again aren't very important, but lower epoch of 5 has lower accuracy
- lower layers have lower accuracy
- the lower learning rates take too long to converge and hence have lower accuracy values

Models that performed well (expected to have close to 95% accuracy)

- had an adam optimizer mostly
- have the higher batch size of 64
- tanh activation has the most positive correlation and importance too
- a higher number of epochs allows for better convergence, hence 10 epochs is better

- a higher learning rate of 0.001 is ideal
- xavier weight initialisation performs better than random initialisation

The observations made from zooming into the parallel coordinate plot are all supported by the correlations and importance graphs too.



▼ Answer 7 (10 Marks)

Parameters of the best model identified are as follows:

activation: tanh, batch size: 64, epochs: 10, hidden layers: 4,
hidden size: 128, learning rate 0.001, optimizer: adam



Run set 2 1 ⋮

▼ Question 8 (5 Marks)

Comparing the performance of two models with the best parameters found earlier, one with loss as MSE and other with loss as cross-entropy.

We notice that the accuracy values on the test and train set are comparable for both the loss functions, the predictions too seem to be comparable going by the confusion matrix. However, we can notice that not much learning seems to be happening with the MSE loss function as the loss curves are fairly constant (though they are lower) as compared to the loss curves of cross-entropy.

☒

Run set 2 1 ⋮



Answer 9 (10 Marks)

link to github code for this assignment


https://github.com/Preethibalamurugan23/DL_Assignment1

▼ Answer 10 (10 Marks)

Three hyper-parameters that we could experiment with for any other dataset would be the optimizer used, activation function and learning rate as the other parameters seem to have lower importance and also their optimum values are common for most datasets and need not be experimented with, higher epochs for exaple are always better.

▼ Self Declaration

I, Preethi - MM21B051, swear on my honour that I have written the code and the report by myself and have not copied it from the internet or other students.

Created with  on Weights & Biases.

<https://wandb.ai/mm21b051-iitmaana/DeepLearning1/reports/Preethi-MM21B051-DA6401-Assignment-1--VmldzoxMTgzODA3OA>