CENG499 - Homework 1

Utku Güngör - 2237477

December 2021

1 Tables

Tables for the different combinations of the hyperparameters(3 layers, 3 different activation functions and 4 different learning rates) with validation accuracy and validation loss are listed below(36 tables in total):

Epoch	Val. Loss
0	4.1199
1	4.0251
2	4.0368
3	3.3913
4	2.7653
5	2.7433
6	2.7826
7	2.8229
8	2.9439
9	2.8499
10	2.9365
11	2.6881
12	2.7428
13	2.7647
14	2.8060
15	3.2148
16	2.7467
17	3.0061
18	2.7805
19	2.8496
Val. Accuracy: **21.320%**	

1 Layer
Act Function: Hardswish
Learning Rate: 10^{-2}

Epoch	Val. Loss
0	3.2168
1	3.0277
2	2.8290
3	2.5929
4	2.4037
5	2.3230
6	2.2494
7	2.2255
8	2.2350
9	2.1852
10	2.1452
11	2.1406
12	2.1466
13	2.1630
14	2.1717
15	2.1311
16	2.1173
17	2.1282
18	2.1314
19	2.1271
Val. Accuracy: **27.160%**	

1 Layer
Act Function: Hardswish
Learning Rate: 10^{-3}

Epoch	Val. Loss
0	3.5383
1	3.4278
2	3.3724
3	3.3245
4	3.2904
5	3.2572
6	3.2267
7	3.2038
8	3.1788
9	3.1588
10	3.1399
11	3.1222
12	3.1037
13	3.0879
14	3.0716
15	3.0553
16	3.0390
17	3.0206
18	3.0031
19	2.9853
Val. Accuracy: **29.320%**	

1 Layer
Act Function: Hardswish
Learning Rate: 10^{-4}

Epoch	Val. Loss
0	4.4875
1	4.0498
2	3.8689
3	3.7670
4	3.6998
5	3.6521
6	3.6154
7	3.5866
8	3.5629
9	3.5426
10	3.5247
11	3.5091
12	3.4950
13	3.4822
14	3.4702
15	3.4590
16	3.4494
17	3.4397
18	3.4310
19	3.4227
Val. Accuracy: **28.760%**	

1 Layer
Act Function: Hardswish
Learning Rate: 10^{-5}

Epoch	Val. Loss
0	4.4420
1	4.4020
2	4.3974
3	4.3888
4	4.3899
5	4.3906
6	4.3930
7	4.3896
8	4.3872
9	4.3872
10	4.3945
11	4.3853
12	4.4108
13	4.3842
14	4.3870
15	4.3851
16	4.3851
17	4.3885
18	4.3860
19	4.3841
Val. Accuracy: **12.130%**	

1 Layer
Act Function: Sigmoid
Learning Rate: 10^{-2}

Epoch	Val. Loss
0	4.7073
1	4.5408
2	4.4658
3	4.4286
4	4.4102
5	4.3993
6	4.3932
7	4.3897
8	4.3866
9	4.3856
10	4.3860
11	4.3841
12	4.3847
13	4.3836
14	4.3841
15	4.3845
16	4.3842
17	4.3848
18	4.3843
19	4.3850
Val. Accuracy: **16.910%**	

1 Layer
Act Function: Sigmoid
Learning Rate: 10^{-3}

Epoch	Val. Loss
0	4.9097
1	4.8719
2	4.8488
3	4.8274
4	4.8059
5	4.7828
6	4.7602
7	4.7366
8	4.7147
9	4.6913
10	4.6689
11	4.6487
12	4.6285
13	4.6098
14	4.5922
15	4.5761
16	4.5605
17	4.5462
18	4.5331
19	4.5206
Val. Accuracy: **22.300%**	

1 Layer
Act Function: Sigmoid
Learning Rate: 10^{-4}

Epoch	Val. Loss
0	5.1087
1	5.0295
2	4.9889
3	4.9635
4	4.9458
5	4.9326
6	4.9223
7	4.9139
8	4.9069
9	4.9008
10	4.8955
11	4.8908
12	4.8865
13	4.8826
14	4.8789
15	4.8756
16	4.8724
17	4.8694
18	4.8665
19	4.8638
Val. Accuracy: **21.790%**	

1 Layer
Act Function: Sigmoid
Learning Rate: 10^{-5}

Epoch	Val. Loss
0	3.1996
1	2.8653
2	2.6891
3	2.6778
4	2.8993
5	2.7732
6	2.8681
7	2.6584
8	2.7163
9	2.7684
10	2.7555
11	2.7935
12	2.6887
13	2.8774
14	2.8597
15	2.6520
16	2.6279
17	2.6813
18	2.8299
19	2.7287
Val. Accuracy: **20.280%**	

1 Layer
Act Function: Relu
Learning Rate: 10^{-2}

Epoch	Val. Loss
0	3.2488
1	3.0646
2	2.8511
3	2.6066
4	2.4497
5	2.3499
6	2.2845
7	2.2436
8	2.2111
9	2.1847
10	2.1655
11	2.1836
12	2.1570
13	2.1601
14	2.1476
15	2.1311
16	2.1314
17	2.1184
18	2.1140
19	2.1128
Val. Accuracy: **26.390%**	

1 Layer
Act Function: Relu
Learning Rate: 10^{-3}

Epoch	Val. Loss
0	3.4951
1	3.4000
2	3.3510
3	3.3241
4	3.2995
5	3.2803
6	3.2610
7	3.2446
8	3.2254
9	3.2083
10	3.1912
11	3.1743
12	3.1570
13	3.1381
14	3.1207
15	3.1051
16	3.0851
17	3.0657
18	3.0446
19	3.0275
Val. Accuracy: **29.700%**	

1 Layer
Act Function: Relu
Learning Rate: 10^{-4}

Epoch	Val. Loss
0	4.5898
1	4.2162
2	4.0672
3	3.9768
4	3.9048
5	3.8411
6	3.7855
7	3.7349
8	3.6876
9	3.6416
10	3.5998
11	3.5544
12	3.5147
13	3.4861
14	3.4641
15	3.4463
16	3.4310
17	3.4190
18	3.4088
19	3.4002
Val. Accuracy: **27.410%**	

1 Layer
Act Function: Relu
Learning Rate: 10^{-5}

Epoch	Val. Loss
0	3.3090
1	3.2865
2	3.2312
3	3.3432
4	3.3150
5	3.3593
6	3.3751
7	3.3303
8	3.3598
9	3.4144
10	3.3550
11	3.3257
12	3.3885
13	3.3205
14	3.3576
15	3.3580
16	3.3807
17	3.3914
18	3.3079
19	3.3770
Val. Accuracy: **21.690%**	

2 Layer
Act Function: Hardswish
Learning Rate: 10^{-2}

Epoch	Val. Loss
0	1.7833
1	1.6662
2	1.6547
3	1.6666
4	1.6593
5	1.6670
6	1.6927
7	1.6847
8	1.7338
9	1.7450
10	1.7248
11	1.7915
12	1.7897
13	1.8106
14	1.8483
15	1.8933
16	1.9183
17	1.9815
18	1.9604
19	2.0450
Val. Accuracy: **42.280%**	

2 Layer
Act Function: Hardswish
Learning Rate: 10^{-3}

Epoch	Val. Loss
0	2.3334
1	1.9745
2	1.8815
3	1.8412
4	1.8038
5	1.7710
6	1.7509
7	1.7353
8	1.7196
9	1.7104
10	1.6965
11	1.6817
12	1.6828
13	1.6691
14	1.6679
15	1.6564
16	1.6599
17	1.6527
18	1.6507
19	1.6403
Val. Accuracy: **43.540%**	

2 Layer
Act Function: Hardswish
Learning Rate: 10^{-4}

Epoch	Val. Loss
0	3.7546
1	3.2845
2	3.0639
3	2.9238
4	2.8063
5	2.7183
6	2.6540
7	2.6005
8	2.5527
9	2.5092
10	2.4667
11	2.4217
12	2.3662
13	2.2796
14	2.2238
15	2.1849
16	2.1516
17	2.1186
18	2.0836
19	2.0440
Val. Accuracy: **34.470%**	

2 Layer
Act Function: Hardswish
Learning Rate: 10^{-5}

Epoch	Val. Loss
0	4.1191
1	4.1191
2	4.1191
3	4.1191
4	4.1191
5	4.1191
6	4.1191
7	4.1191
8	4.1191
9	4.1191
10	4.1191
11	4.1191
12	4.1191
13	4.1191
14	4.1191
15	4.1191
16	4.1191
17	4.1191
18	4.1191
19	4.1191
Val. Accuracy: **10.010%**	

2 Layer
Act Function: Sigmoid
Learning Rate: 10^{-2}

Epoch	Val. Loss
0	4.1195
1	4.1191
2	4.1187
3	4.1188
4	4.1186
5	4.1183
6	4.1180
7	4.1181
8	4.1182
9	4.1181
10	4.1186
11	4.1182
12	4.1186
13	4.1188
14	4.1191
15	4.1198
16	4.1195
17	4.1194
18	4.1200
19	4.1198
Val. Accuracy: **10.120%**	

2 Layer
Act Function: Sigmoid
Learning Rate: 10^{-3}

Epoch	Val. Loss
0	4.1348
1	4.1233
2	4.1208
3	4.1199
4	4.1194
5	4.1191
6	4.1189
7	4.1189
8	4.1188
9	4.1188
10	4.1188
11	4.1188
12	4.1188
13	4.1188
14	4.1187
15	4.1187
16	4.1186
17	4.1186
18	4.1186
19	4.1185
Val. Accuracy: **10.940%**	

2 Layer
Act Function: Sigmoid
Learning Rate: 10^{-4}

Epoch	Val. Loss
0	4.5189
1	4.3047
2	4.2170
3	4.1736
4	4.1504
5	4.1374
6	4.1300
7	4.1256
8	4.1230
9	4.1214
10	4.1205
11	4.1199
12	4.1196
13	4.1194
14	4.1193
15	4.1192
16	4.1191
17	4.1190
18	4.1190
19	4.1190
Val. Accuracy: **15.030%**	

2 Layer
Act Function: Sigmoid
Learning Rate: 10^{-5}

Epoch	Val. Loss
0	4.1384
1	4.1813
2	4.0973
3	4.1325
4	4.1139
5	4.1072
6	3.9666
7	3.8954
8	3.8910
9	3.8965
10	3.9401
11	3.8805
12	3.9261
13	3.8770
14	3.8567
15	3.8507
16	3.8588
17	3.9189
18	3.8863
19	3.9204
Val. Accuracy: **20.200%**	

2 Layer
Act Function: Relu
Learning Rate: 10^{-2}

Epoch	Val. Loss
0	1.8512
1	1.7309
2	1.7020
3	1.6898
4	1.6811
5	1.6663
6	1.6718
7	1.7046
8	1.6977
9	1.7086
10	1.7266
11	1.7270
12	1.8039
13	1.8185
14	1.8371
15	1.8518
16	1.8547
17	1.8776
18	1.8932
19	1.9399
Val. Accuracy: **42.560%**	

2 Layer
Act Function: Relu
Learning Rate: 10^{-3}

Epoch	Val. Loss
0	2.0267
1	1.9102
2	1.8622
3	1.8248
4	1.7984
5	1.7760
6	1.7553
7	1.7390
8	1.7269
9	1.7154
10	1.7081
11	1.6986
12	1.6903
13	1.6774
14	1.6860
15	1.6705
16	1.6692
17	1.6644
18	1.6536
19	1.6550
Val. Accuracy: **42.570%**	

2 Layer
Act Function: Relu
Learning Rate: 10^{-4}

Epoch	Val. Loss
0	3.4495
1	2.8316
2	2.5226
3	2.3437
4	2.2400
5	2.1711
6	2.1216
7	2.0833
8	2.0536
9	2.0283
10	2.0090
11	1.9907
12	1.9765
13	1.9626
14	1.9515
15	1.9413
16	1.9314
17	1.9236
18	1.9159
19	1.9079
Val. Accuracy: **34.550%**	

2 Layer
Act Function: Relu
Learning Rate: 10^{-5}

Epoch	Val. Loss
0	2.1517
1	2.1686
2	2.1450
3	2.1590
4	2.1138
5	2.1242
6	2.1836
7	2.1667
8	2.1442
9	2.1615
10	2.1753
11	2.1222
12	2.1814
13	2.1487
14	2.1246
15	2.1436
16	2.1639
17	2.1475
18	2.1864
19	2.1805
Val. Accuracy: **16.480%**	

3 Layer
Act Function: Hardswish
Learning Rate: 10^{-2}

Epoch	Val. Loss
0	1.6977
1	1.6621
2	1.6091
3	1.6016
4	1.6104
5	1.6147
6	1.6747
7	1.7109
8	1.7817
9	1.8349
10	1.8678
11	1.9438
12	2.0694
13	2.1157
14	2.2029
15	2.2953
16	2.4044
17	2.4822
18	2.5534
19	2.6457
Val. Accuracy: **41.660%**	

3 Layer
Act Function: Hardswish
Learning Rate: 10^{-3}

Epoch	Val. Loss
0	1.8985
1	1.8221
2	1.7690
3	1.7380
4	1.6999
5	1.6749
6	1.6581
7	1.6366
8	1.6197
9	1.6143
10	1.6055
11	1.5930
12	1.5902
13	1.5834
14	1.5856
15	1.5836
16	1.5819
17	1.5924
18	1.5777
19	1.5803
Val. Accuracy: **45.320%**	

3 Layer
Act Function: Hardswish
Learning Rate: 10^{-4}

Epoch	Val. Loss
0	2.1381
1	2.0746
2	2.0337
3	2.0038
4	1.9787
5	1.9577
6	1.9402
7	1.9248
8	1.9119
9	1.8999
10	1.8898
11	1.8807
12	1.8711
13	1.8634
14	1.8555
15	1.8481
16	1.8411
17	1.8344
18	1.8282
19	1.8225
Val. Accuracy: **35.870%**	

3 Layer
Act Function: Hardswish
Learning Rate: 10^{-5}

Epoch	Val. Loss
0	1.9423
1	1.9456
2	1.9530
3	1.9570
4	1.9311
5	1.9005
6	1.9075
7	1.9022
8	1.9194
9	1.9162
10	1.8967
11	1.8759
12	1.8986
13	1.8822
14	1.9024
15	1.8896
16	1.8788
17	1.8871
18	1.9160
19	1.8970
Val. Accuracy: **33.380%**	

3 Layer
Act Function: Sigmoid
Learning Rate: 10^{-2}

Epoch	Val. Loss
0	1.9526
1	1.8575
2	1.7920
3	1.7602
4	1.7423
5	1.7215
6	1.7079
7	1.7067
8	1.6911
9	1.7085
10	1.7182
11	1.7498
12	1.7402
13	1.7779
14	1.7864
15	1.8127
16	1.8476
17	1.8827
18	1.9401
19	1.9625
Val. Accuracy: **40.070%**	

3 Layer
Act Function: Sigmoid
Learning Rate: 10^{-3}

Epoch	Val. Loss
0	2.0825
1	2.0304
2	2.0081
3	1.9915
4	1.9762
5	1.9654
6	1.9551
7	1.9401
8	1.9284
9	1.9167
10	1.9068
11	1.8955
12	1.8866
13	1.8753
14	1.8676
15	1.8555
16	1.8468
17	1.8388
18	1.8323
19	1.8220
Val. Accuracy: **36.280%**	

3 Layer
Act Function: Sigmoid
Learning Rate: 10^{-4}

Epoch	Val. Loss
0	2.2696
1	2.2384
2	2.2096
3	2.1844
4	2.1621
5	2.1428
6	2.1258
7	2.1110
8	2.0982
9	2.0875
10	2.0783
11	2.0706
12	2.0641
13	2.0585
14	2.0533
15	2.0492
16	2.0454
17	2.0414
18	2.0380
19	2.0351
Val. Accuracy: **28.470%**	

3 Layer
Act Function: Sigmoid
Learning Rate: 10^{-5}

Epoch	Val. Loss
0	2.1295
1	2.1213
2	2.0883
3	2.1491
4	2.1435
5	2.1216
6	2.1146
7	2.1082
8	2.0938
9	2.0954
10	2.1857
11	2.1792
12	2.1233
13	2.1383
14	2.1287
15	2.1206
16	2.1063
17	2.1720
18	2.1665
19	2.1388
Val. Accuracy: **18.500%**	

3 Layer
Act Function: Relu
Learning Rate: 10^{-2}

Epoch	Val. Loss
0	1.7546
1	1.6821
2	1.6634
3	1.6211
4	1.6190
5	1.6501
6	1.6449
7	1.6492
8	1.6908
9	1.7462
10	1.7832
11	1.8194
12	1.8945
13	1.9420
14	1.9898
15	2.0561
16	2.1102
17	2.2375
18	2.2823
19	2.3571
Val. Accuracy: **42.320%**	

3 Layer
Act Function: Relu
Learning Rate: 10^{-3}

Epoch	Val. Loss
0	1.8810
1	1.8025
2	1.7523
3	1.7137
4	1.6789
5	1.6606
6	1.6444
7	1.6268
8	1.6189
9	1.6072
10	1.6112
11	1.5980
12	1.6032
13	1.6056
14	1.5949
15	1.6151
16	1.6139
17	1.6125
18	1.6178
19	1.6149
Val. Accuracy: **45.370%**	

3 Layer
Act Function: Relu
Learning Rate: 10^{-4}

Epoch	Val. Loss
0	2.1328
1	2.0513
2	2.0023
3	1.9699
4	1.9454
5	1.9260
6	1.9084
7	1.8943
8	1.8815
9	1.8697
10	1.8601
11	1.8500
12	1.8421
13	1.8347
14	1.8271
15	1.8201
16	1.8133
17	1.8067
18	1.8013
19	1.7956
Val. Accuracy: **36.700%**	

3 Layer
Act Function: Relu
Learning Rate: 10^{-5}

2 Validation Set

- Validation sets are created just like in the sample code. Used method is 5-fold validation. 80% of the set is used for training whereas the remaining 20% for validation. *random_split* function is applied to break the set into two parts.

3 Sanity Check

3.1 Loss

We can use Cross Entropy loss function to compute loss:

$$H(p,q) = -\sum_{x \in X} p(x) log(q(x))$$

$$p(x) = 1 and since we have 10 labels, q(x) = 0.1$$

$$H(p,q) = -\sum_{x \in X} 1.log(0.1)$$

$$H(p,q) \approx 2.30$$

When I compute the validation loss before training the data, I reached 2.3061 as the validation loss which is close to what I expected.

3.2 Accuracy

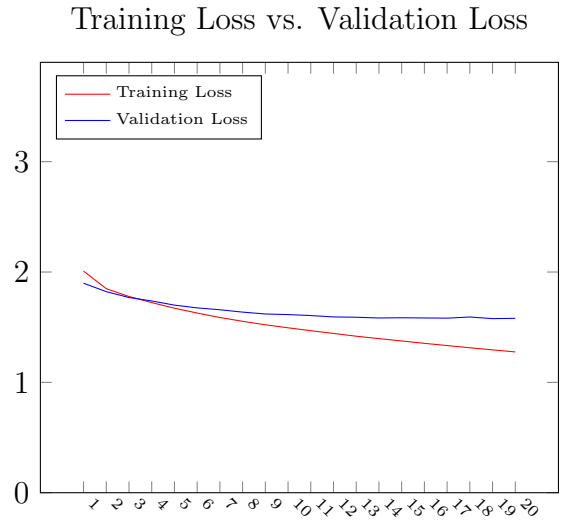
Since there are 10 categories, each will be expected to have 0.1 probability. Before training the data, I computed the expected test accuracy as 10.280 which is again close to my expectation.%.

4 Test Accuracies and The Best Performing Hyperparameter Configurations

- At first, I tried learning rates in range $(10^{-1}, 10^{-4})$. But for 10^{-1}, test accuracy was about 10-15% for most of the cases and it was not affected from the activation function and layer numbers in a positive way. Then, I changed the range to $(10^{-2}, 10^{-5})$ which offer better results. In addition, used epoch number during the model training is 20.

- As expected, test accuracies were near 30% for 1 layer and near 40% for 2 layer. For 3 layer, it would be a little more but it is still not too low, around 45%. It would result in better accuracies if it the epoch number was higher but that would increase the required time.

- The best performing hyperparameter configurations use a 3-layer network, Hardswish as the activation function and 10^{-4} as the learning rate. The test accuracy is 45.160% for this configuration. Its graph with the related parameters is below:



The training loss and the validation loss are quite close to each other in this configuration. Towards the end, validation losses get more stable and if I keep training, maybe it would go much higher and cause overfitting which shows stopping at that point is very crucial to achieve higher test accuracy. Therefore, that configuration gives the best result with its test accuracy, relationship between the training loss and the validation loss, and the training time.

5 Detecting and Avoiding Overfitting

- Sudden changes in validation or training loss can cause overfitting. It is basically when validation loss is much higher than training loss. So, if validation loss starts to increase or training loss starts to decrease considerably, then we can understand that there is a potential overfitting.

- The gap between the training loss and validation loss is not big. That

shows there is no overfitting or underfitting, so our model does not memorize the data as desired. As I observe, for most of the configurations, learning rate 10^{-2} results in overfitting. Its hard to jump from local minimas when we have smaller step size, I guess this is the reason behind overfitting with that learning rate. 10^{-4} gives optimal results in many cases, so that is the strategy I applied to avoid overfitting. K-fold cross validation also helped me by splitting the data and if the difference I mentioned above is considerable, then I know I need to choose simpler models to prevent overfitting.

6 Measuring the Performance

Accuracy can be considered as a suitable metric to measure the performance of this network since test accuracies were near required values for this homework. In addition to the hyperparameter configurations, changes in the batch size or epoch number can have some positive or negative effects according to the values chosen.

7 Learning Rate and Batch Size

7.1 Learning Rate

- One of the hyperparameters I changed is the learning rate. As mentioned before, small learning rate such as 10^{-1} results in bad test accuracies and high validation and training losses whereas higher learning rates result in different values for those variables in a positive way.

- For the learning rate 10^{-1}, loss values suddenly start to increase in some point which shows smaller rates can cause overfitting. This sitution will probably happen for too high learning rates. Therefore, finding an optimal learning rate is important to prevent overfitting.

- Small learning rates can get stuck on processes while large learning rates can help jump and get rid of those situations.

- Large learning rates allow model to learn faster while small learning rates can yield slow processes.

7.2 Batch Size

The batch size I usually used was 32. I also tried with 4 and 512, and observed the changes as following:

- For the small batch size, there is a small decrease on training and validation losses.

- For the small batch size, there is an increase on training and validation losses.

- Small batch size give better test accuracy whereas large batch size cause lower test accuracies.

- For a fixed number of epochs(20 in our case), larger batch sizes take fewer steps and hence they do not generalize well.

- Smaller batch sizes can converge faster than large batch sizes, however, a large batch size can reach optimum minima that a small batch size cannot reach.