Experiment Zero Report

Yue Li

September 5, 2021

Experiment description

Experiment number:0

Purpose

Hypothetically, the difficulty of the samples perceived by the model is proportional to the percentage of noise present in the sample, therefore a simple curriculum can be that order the samples by the percentage of noise present in the image in ascending order and feed them into the model. The expectation is that the end accuracy of the model in the experiment group on the test set will be better than that of the controlled group. Given that the images in the MINIST dataset have no noises, to make this experiment work, artificial noise which is pepper noise will be added to the samples.

Simple curriculum

This simple curriculum adds artificial pepper noise to all MINIST samples. The experiment divides training into two separate groups.

Group 1: with noise, X percent fixed noise (controlled group or non-curriculum group with varying noise ranging from 10% to 40%).
Group 2: with noise, noise increases by 1.5% in each epoch stating from 0(experiment group or curriculum group where the corresponding X is the max noise percent).

Training strategy

For group 1: do nothing in the training loop. For group 2: increase percentage of pepper noise by 1.5% in each iteration.

Experiment parameters

Dataset: MINIST
Loss function: cross entropy loss
Learning rate:0.003
Optimizer: adaptive momentum

Epoch:20
Noise
Artificial Pepper Noise
Noise percent:

- 1: X percent fixed
- 2: staring from 0%, incrementing by 1.5% in each epoch

Runs

Total number of runs of group 1: 10
Total number of runs of group 2: 10

Tools

Programming language: Python
Deep learning framework: Pytorch

Model

Residual network with 2 resblock and ReLU as non-linearity function, a graphic representation of the model:

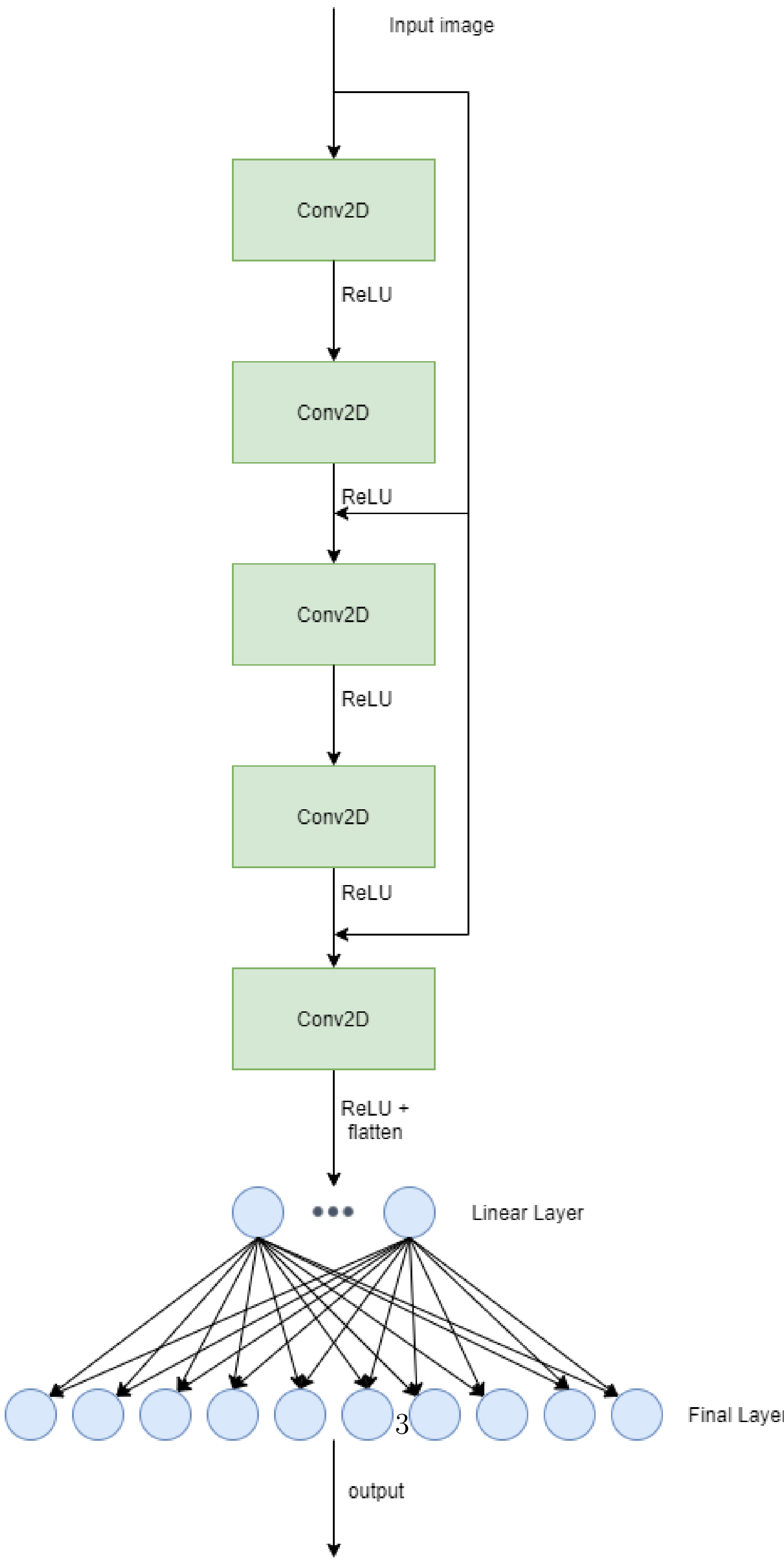

Figure 1: ResNet architecture used in this experiment

Computing Environment
Windows 10 with i7 6500U processor (for noise percent 10-19 and 30-40)
Windows 10 with Ryzen 5 processor (for noise percent 20-29)

Result Analysis

The following are the summary table of the sequences executed with different level of noises:

Table 1: summary of sequence summary with curriculum

Noise Percent	Best Train Accuracy	Mean train Accuracy	Best Test Accuracy	Mean Test Accuracy	Mean Validation Accuracy
10.0	0.9644	0.9149	0.8653	0.8355	0.8655
11.0	0.958	0.9043	0.8576	0.8265	0.8471
12.0	0.9532	0.9007	0.8576	0.828	0.859
13.0	0.95	0.8989	0.8635	0.8312	0.8628
14.0	0.9499	0.895	0.8677	0.8353	0.8627
15.0	0.9464	0.891	0.8611	0.8318	0.8631
16.0	0.9446	0.8834	0.8681	0.8292	0.8602
17.0	0.9351	0.8772	0.8675	0.8281	0.8684
18.0	0.9318	0.8743	0.8642	0.8294	0.863
19.0	0.9297	0.8697	0.8629	0.8278	0.8648
20.0	0.9302	0.8726	0.8638	0.8326	0.8632
21.0	0.931	0.8689	0.8657	0.8294	0.8599
22.0	0.925	0.864	0.8632	0.8272	0.8576
23.0	0.9183	0.8531	0.8592	0.8215	0.853
24.0	0.9206	0.8548	0.8606	0.8231	0.8626
25.0	0.916	0.8491	0.8644	0.8248	0.8593
26.0	0.9128	0.8424	0.8674	0.8291	0.8614
27.0	0.9101	0.8414	0.8709	0.8314	0.8678
28.0	0.908	0.8405	0.8706	0.8329	0.8698
29.0	0.9094	0.8337	0.8582	0.8198	0.8602
30.0	0.8965	0.828	0.8641	0.8253	0.8629
31.0	0.8974	0.8295	0.8646	0.8232	0.8694
32.0	0.8943	0.8231	0.8585	0.8227	0.8453
33.0	0.8943	0.8232	0.8645	0.8223	0.8651
34.0	0.8813	0.8121	0.8683	0.8247	0.8704
35.0	0.8826	0.8072	0.8638	0.8227	0.8652
36.0	0.8774	0.7999	0.866	0.8228	0.8619
37.0	0.8759	0.8017	0.8611	0.8184	0.8582
38.0	0.8717	0.7971	0.8695	0.8232	0.8694
39.0	0.8816	0.8015	0.859	0.8201	0.8568
40.0	0.8802	0.8024	0.8604	0.8197	0.8459

The starting point of analysis can be the Mean Loss plot as given below:

Table 2: summary of sequence summary without curriculum

Noise Percent	Best Train Accuracy	Mean train Accuracy	Best Test Accuracy	Mean Test Accuracy	Mean Validation Accuracy
10.0	0.9796	0.7822	0.6693	0.5743	0.6873
11.0	0.9217	0.7307	0.5886	0.4972	0.5985
12.0	0.9636	0.7144	0.614	0.4886	0.6335
13.0	0.951	0.6868	0.5561	0.4521	0.5684
14.0	0.9151	0.6921	0.5521	0.4575	0.564
15.0	0.9125	0.6747	0.4898	0.3991	0.4956
16.0	0.9087	0.6529	0.4973	0.407	0.5017
17.0	0.9102	0.6558	0.4673	0.3886	0.4805
18.0	0.9137	0.6061	0.4323	0.3422	0.4436
19.0	0.9508	0.5967	0.4527	0.3339	0.4685
20.0	0.8863	0.5922	0.4433	0.3463	0.4536
21.0	0.8285	0.5339	0.4036	0.3179	0.4152
22.0	0.8391	0.5357	0.3807	0.2992	0.3913
23.0	0.8423	0.5275	0.3773	0.2891	0.3844
24.0	0.8574	0.5227	0.3749	0.2918	0.3874
25.0	0.7881	0.4667	0.3441	0.2714	0.3505
26.0	0.7798	0.4507	0.3063	0.2335	0.31
27.0	0.7586	0.4635	0.3168	0.2487	0.3205
28.0	0.7479	0.4356	0.3372	0.2411	0.3335
29.0	0.8011	0.4698	0.3243	0.2487	0.329
30.0	0.6735	0.3745	0.2783	0.2147	0.2812
31.0	0.7731	0.464	0.3275	0.2396	0.3292
32.0	0.6616	0.3617	0.2863	0.2105	0.2844
33.0	0.6546	0.3664	0.2902	0.2159	0.2837
34.0	0.6693	0.3744	0.2655	0.2039	0.2687
35.0	0.7185	0.4005	0.3147	0.2207	0.3163
36.0	0.6298	0.3777	0.2674	0.2035	0.2562
37.0	0.6626	0.3865	0.2679	0.2076	0.2717
38.0	0.528	0.3054	0.2197	0.1763	0.2139
39.0	0.5827	0.3393	0.2597	0.1959	0.2669
40.0	0.6384	0.3571	0.2747	0.2	0.2782

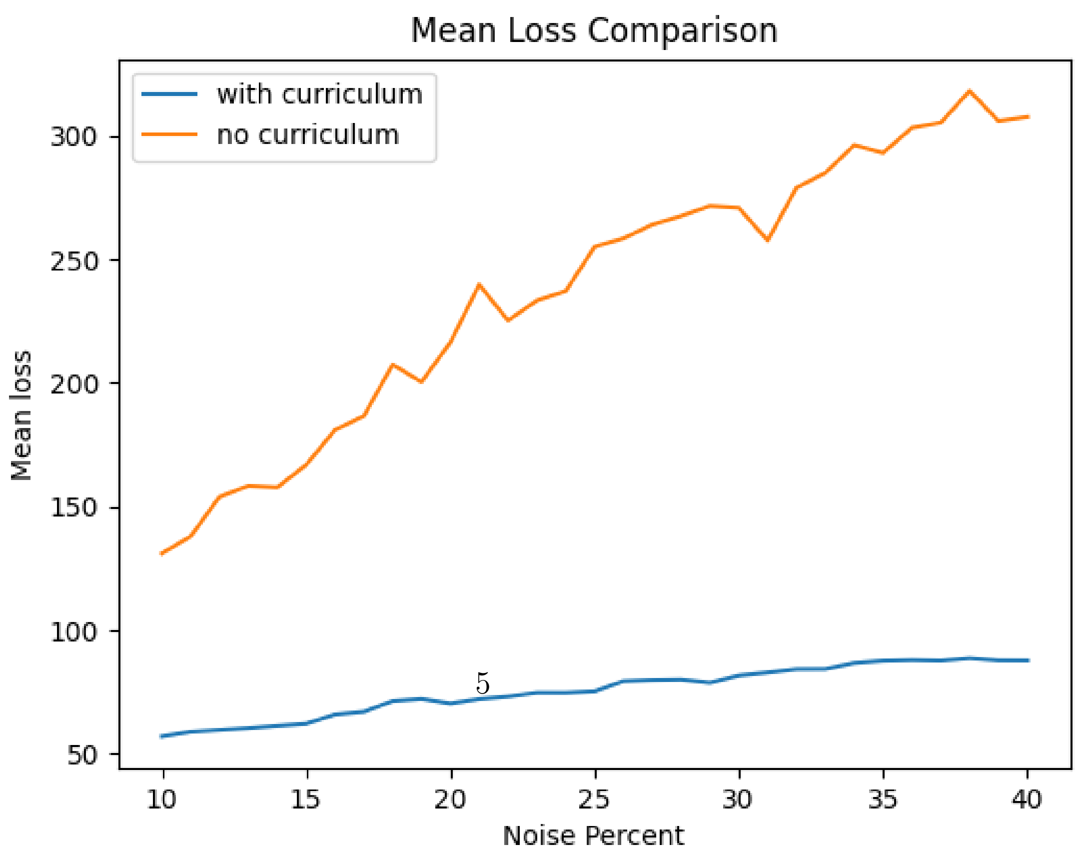

Figure 2: Loss of the curriculum group and the non-curriculum group

The horizontal axis represents the percent of noise applied to the images. The vertical axis represents the mean loss which is obtained by taking the mean of the mean of loss of each iteration of a particular run for all runs. As clearly indicated by the graph. The loss of the non-curriculum group increases faster than the curriculum group as the percent of noise increases. This provides initial evidence that using curriculum can reduce loss.

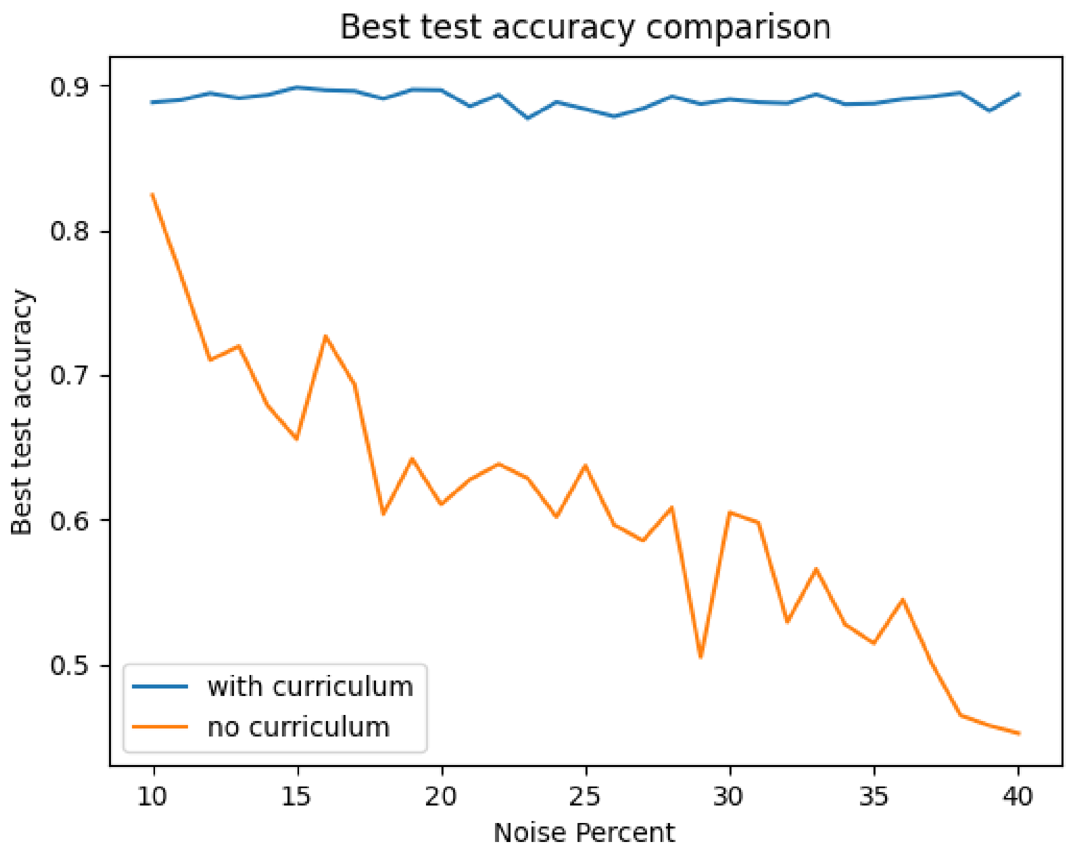

Figure 3: Best test accuracy comparison

The above figure of best test accuracy examines the effect on the peek performance a model can achieve when using curriculum as well as not using curriculum. The best test accuracy of one run under a particular noise percent is obtained by taking the max test accuracy value across all iterations. Similarly, the best accuracy amongst runs is acquired using the max value from the vector $\vec{a}$ that stores the best test accuracy of each run $a_{best} = max(\vec{a})$ As shown by the above figure titled Best test accuracy comparison. The best test accuracies from the curriculum group are always better then the non-curriculum group.

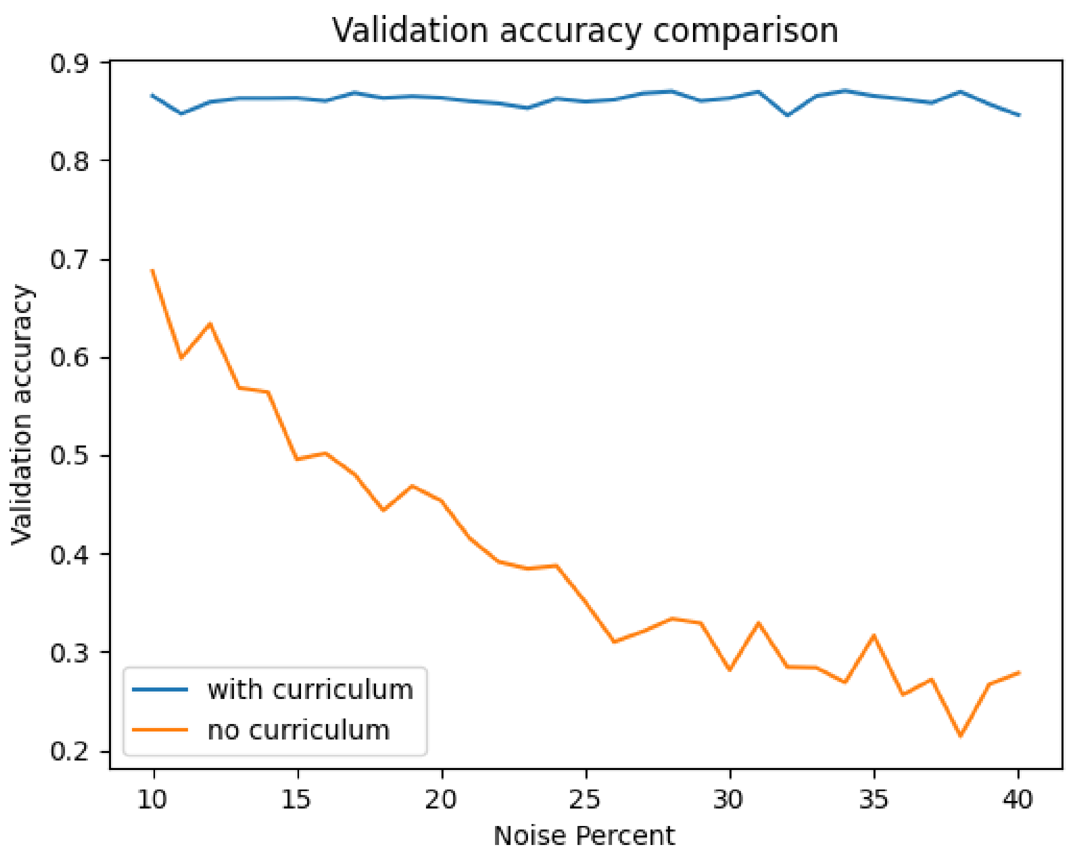

Figure 4: Validation accuracy of the curriculum group and the non-curriculum group

The noise percent vs validation accuracy graph for noise ranging from 10% to 40% is given above. The validation accuracy is the mean validation accuracy of 10 runs for each noise percent. As clearly indicated by this figure.Curriculum group's performance is very stable on the validation dataset with a mean accuracy of 0.86, whereas the non-curriculum group's performance on the validation dataset decreases approximately quadradically as the noise applied to the images increases.

The following calculations attempt to give a quantitative analysis regarding how much performance gain can be acquired on the validation dataset when using curriculum.

Denote the mean validation accuracy of each run in the curriculum group and the non-curriculum group as A_i, B_i respectively, define the pair wise difference $D_i = A_i - B_i$, then the mean of the difference of validation accuracy is:

$$\mu_d = \frac{\sum_{i=0}^{n} D_i}{n}$$

From the sequence summary results, average validation performance increase $\mu_d = 0.45$ Ideally, experiment 0 should be conducted multiple times in order to get a confidence interval for μ_d.

Conclusion

Based on the experimental data, it's very likely that curriculum can provide huge performance boost to ResNet. An average of 45% increase in performance is promising—although it's just one experiment.To investigate whether curriculum learning is a general methodology, other models should be examined as well. Also different curriculum can impact performance as well. Therefore, different curriculums should be investigated as well.