Dharmit Prajapati
1219597132

Fundamentals of Statistical Learning Project3 Report

In this project we have to classify a CIFAR-10 datatest using convolution neural network. The aim of the project is to learn how the changes in hyperparameters impacts the accuracy of the model.

We are given a baseline code and we performed hyperparameter tuning to record the impact. Below are the results:

Baseline Accuracy: 82.58000016212463 , Baseline loss : 0.6287005543708801

No.	Hyperparameter	Baseline value	New value	Accuracy	Loss
1a)	Learning rate	0.01	0.1	78.25	6.67714834213256
1b)	Learning rate	0.01	0.001	86.58	0.50549238920211
2	Kernel size of first layer	3X3	5x5	84.45	0.54670417308807
3	Learning rate optimizer	Adam	SGD	83.71	0.51554197072982
4	Batch normalization	Present	Removed	10.00	2.30360317230224
5	Dropout layer	Present	Removed	78.94	4.13423871994018
6a)	Batch size	64	32	84.58	0.54108190536499
6b)	Batch size	64	128	86.34	0.58072316646575

1)**Change in the learning rate:**
Learning rate indicates the response of the updation of model towards estimated error. If the error rate is too small the training process will be long, if it too large the updated/new weights might be suboptimal.

0.1 learning rate is too large and has resulted in lesser accuracy as compared to accuracy achieved with 0.001 and 0.01 learning rate.

2)**Change in the kernel size of first layer:**

Increasing the kernel size indicates it has more parameters to train which might give better accuracy(in our case it does increase the accuracy when the kernel size is made 5x5 from 3x3) but that comes at a cost of increased computational complexity. This is evident by the increase in accuracy achieved for 5x5 as compared to 3x3 kernel in the first layer. But increase in the kernel size doesn't necessarily mean the accuracy will increase. Bigger kernel size may miss out on features collected by 3x3 as compared to 5x5 as 3x3 will detect them better.

3) **Change optimizer to SGD:**

SGD generalizes better than Adam, even though Adam converges faster than SGD. Since generalization is better overall, the performance on the test data is better as compared to baseline, which is supported by the accuracy results.

4) **Remove all the batch normalization layers in the network:**

Normalization helps in scaling of inputs which makes sure all of them fall within a similar range, which is important as the inputs might be from different sources and will have different units of measurements. Not normalizing the data makes the training process harder and the learning speed also decreases.

Batch Normalization is a normalization technique performed within Neural network layers rather than on the raw data. It is done on batches rather than on the whole training set resulting in faster training and higher learning rates. This provides regularization and helps avoid overfitting. Removing batch normalization has the most drastic effect on the accuracy which is evident from the results, indicating the model is overfitted on training data and can't perform well on test data.

5) **Remove all the dropout layers in the network:**

Dropout layer drops out some neurons randomly during training at a rate mentioned in the argument. Dropping out a neuron means that it will no longer participate in the further training process and the other neurons have to take up the responsibility of generalization and make predictions. As this is done randomly in each iteration, neurons are not assigned specific weights making it more generalized hence avoiding overfitting.

The decrease in the accuracy after removing the dropout layers indicates that the model might have overfitted a bit.

6) **Change in batch size**

Batch size indicates the number of samples that will propagated throughout the network for learning in each pass/iteration.
Small batch size may lead to sub-optimal learning as only a small portion of the whole data is involved in the learning process for that pass, the model may not see a variety of data if the batch size is small.
If it's too large, the training and learning process will take more time and memory.
That's why the accuracy of 128 batch size is better than that of 32 and 64(baseline).