A study of effect of pruning in a fully connected layer of a CNN architecture used for classification.

Afsar Ahamed Asaraf Ali[1][0000−0003−3511−7463]

[1] Research School of Computer Science
Australian National University, Canberra Australia
u7099664@anu.edu.au

Abstract. This paper discusses how pruning or removal of redundant neurons in a fully connected layer in a CNN architecture affects performance of that CNN model in an image classification problem.The dataset used for training and testing the model is a part of vehicle-x dataset that corresponds to the first 20 vehicleIDs of 1362 vehicleIDs, where vehicleID is one of the attributes of an image in the dataset and is required to predicted by the model by inputting the image.The baseline CNN architecture used for classifying images is LeNet-5 Architecture published by LeCun et al., in 1998.The pruning technique is implemented in the penultimate fully connected layer(sixth layer) of LeNet-5 Architecture.The fully connected layer chosen for pruning consists of 84 neurons in it's baseline model. Pruning of an neuron in the fully connected layer is done one by one (followed by training the model again) and the performance of the model is assessed for every time a neuron is pruned.A neuron that shares maximum similarity with another neuron based on cosine similarity of their respective activation vectors is chosen for pruning at each step.The paper "Simulating Content Consistent Vehicle Datasets with Attribute Descent" [1] that used the same dataset (vehicle-x) does not deal with a classification problem.Therefore, comparison of the results may not be done. However, It would be interesting to compare the results of the pruned model with respected to the baseline model. Testing accuracy of the baseline model is obtained to be 0.57 . Testing accuracy is obtained every time a neuron is pruned followed by retraining of the model.By plotting these test accuracies with respect to corresponding number of neurons that are pruned, It's evident that there is a significant fall in testing accuracy in the model when the number of neurons that are pruned is more than 50. This explains there are approximately 50 redundant neurons in the penultimate fully connected layer of the model.

Keywords: pruning · activation vector · progressive compression

1 Introduction

1.1 Motivation and goal

Despite the great success of Convolutional Neural Networks (CNNs) in various visual recognition tasks, the high computational and storage costs of such deep networks impede their deployments. To this end, considerable attention has been given to the pruning techniques[2].The main motivation of this study is to develop an intuition about how pruning can affect performance of a chosen CNN architecture and to what extent it can be pruned so the performance of the pruned model is not significantly different comparing to the baseline in terms of performance.

1.2 Dataset

A dataset called vehicle-x was used to train the Baseline Image classification model which is LeNet-5 Architecture published by LeCun et al., in 1998. It consists of images of several vehicles pictured at different physical conditions.This dataset was particularly chosen for this problem because the vehicles are generally symmetric and has repetitive features like similar looking tyres,doors etc.., lots of the information in the image can be redundant and repetitive .The dataset is also provided with a set of attributes for each image which includes vehicle ID, vehicle orientation, light intensity, light direction, camera distance,camera height, vehicle type and vehicle color.The description for them is given below :

Vehicle ID : Provides Identification number for a vehicle.
Vehicle orientation: It is the horizontal viewpoint of a vehicle and takes a value between 0°and 359°[1]
Light direction: It describes daylight as cars are generally presented in outdoor scenes.[1]
Light intensity : It represents degree of brightness of the light.
Camera height : It describes the vertical distance from the ground.[2]
Camera distance : It determines the horizontal distance from vehicles [1].

For this CNN classification problem, a part of this dataset corresponding to first 10 vehicle ID out of 1362 is considered for labelling. The chosen 10 vehicle ID for labelling are numbers ranging from 1 to 10.This is done so

because more number of classes in this dataset posed to be challenging in terms of achieving good accuracy and eventually studying the pruning effect.The chosen label Vehicle ID is very well distributed.The histogram plots of labels of the entire dataset,training set,testing set, validation set is shown in Fig. 1.Data are well distributed and balanced among the chosen 10 classes in the entire chose dataset. Before using the images for training the model, necessary pre-processing is done to the images. Initially ,the images are converted to grayscale so that it can implemented by the chosen model and basic preprocessing of standardization is performed to ensure each corresponding pixel has a similar data distribution.Since, the dataset consists of images of vehicles, it's reasonable to expect that most of the images have close resemblance due to basic structure of any vehicle. In order to deal with this, random horizontal flip,random rotation , random changes in the brightness, contrast and saturation of an image are implemented to improve the variations in the images.Finally, the image is resized to 32x32 so that it can be used in the chosen standard CNN architecture , which is LeNet-5 Architecture published by LeCun et al., in 1998.

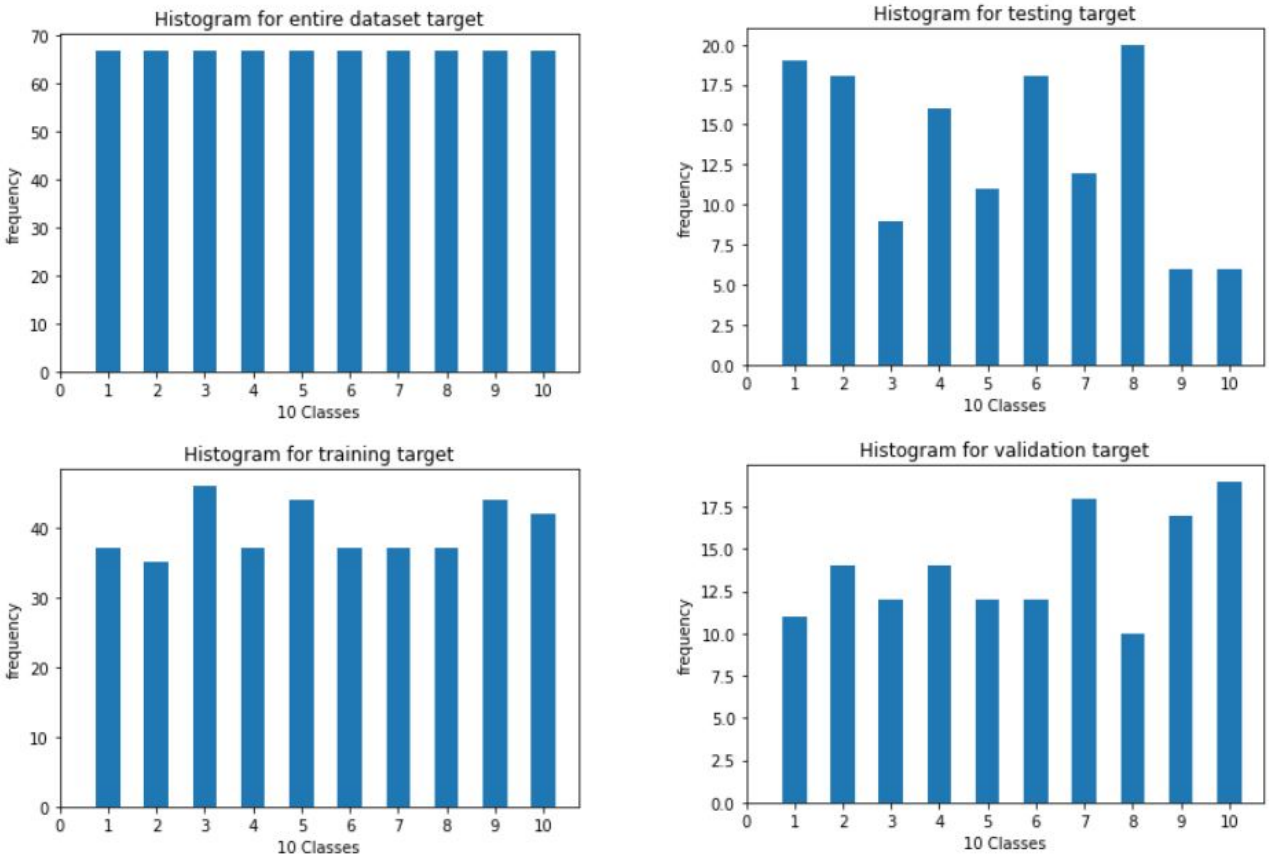

Fig. 1. Histogram plots for entire, training, testing and validation dataset.

1.3 Hyper parameter and packages used

Important hyper parameters used in this model are batch size,learning and number of epochs.The model used to study the effect of pruning is pre-decided. which is LeNet-5 Architecture published by LeCun et al., in 1998.Therefore, studying the effect of pruning is prioritized. Best validation accuracy for different learning rates is given in Table 1 below.

Table 1. Best validation accuracy at different learning rate

Learning rate	Accuracy(in percentage)
0.1	43.81
0.01	73.0 5
0.05	61.54
0.001	21.82

It's evident that learning rate = 0.01 is the most optimal learning rate.Batch size is chosen to be 16 to deal with GPU limitation of the system used for computing the model and batch size is not chosen to be any less than 16 so that the model can converge faster.Number of epochs is chosen to be 200, as validation accuracy converges before 160th epoch for baseline and for pruned model with different number of pruned neurons.

The main packages used for implementing this model and the pruning techniques are numpy,torch and pillow. Kaggle website is used for the GPU service to run deep-learning algorithm.

1.4 Model

The model used for the image classification problem following by the study of affect of pruning is pre-decided, which is LeNet-5 Architecture published by LeCun et al., in 1998.The architecture is straightforward and simple to understand that's why it is mostly used as a first step for teaching Convolutional Neural Network[3] . Therefore, this is used in this classification problem , where the primary goal is to study the effect of pruning in the performance of a CNN classification model.

The LeNet-5 architecture consists of two sets of convolutional and average pooling layers, followed by a flattening convolutional layer, then two fully-connected layers and finally a softmax classifier[3] as shown in Fig. 2 [4]. They are presented below :

First Layer: The input for LeNet-5 is a 32×32 grayscale image which passes through the first convolutional layer with 6 feature maps or filters having size 5×5 and a stride of one. The image dimensions changes from 32x32x1 to 28x28x6.

Second Layer: The Second layer is a average pooling layer or sub-sampling layer with a filter size 2×2 and a stride of two. The resulting image dimensions will be reduced to 14x14x6.

Third Layer: The Third layer is a second convolutional layer with 16 feature maps having size 5×5 and a stride of 1.

Fourth Layer: The fourth layer is again an average pooling layer with filter size 2×2 and a stride of 2. This layer is the same as the second layer except it has 16 feature maps so the output will be reduced to 5x5x16.

Fifth Layer: The fifth layer is a fully connected convolutional layer with 120 feature maps each of size 1×1. Each of the 120 units in fifth layer is connected to all the 400 nodes (5x5x16) in the fourth layer.

Sixth Layer: The sixth layer is a fully connected layer with 84 units.

Seventh Layer: Seventh layer is a fully connected softmax output layer with 10 possible values corresponding to the digits from 0 to 9.[3]

hyperbolic tangent function (Tanh) is used as activation function after every convolution layer.Sigmoid function is used after the last layer.

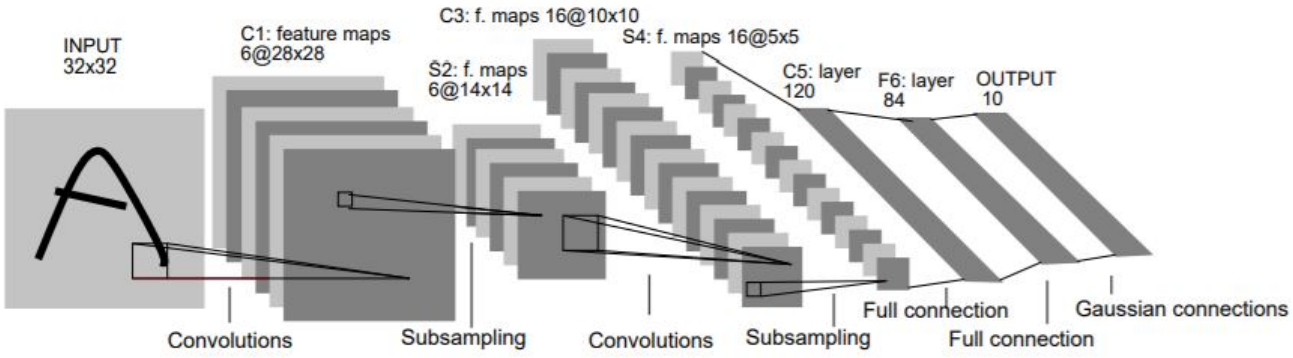

Fig. 2. LeNet-5 Architecture

1.5 Pruning

In a fully connected layer, pruning is a process of removal of neurons based on redundancy.A neuron in a fully connected layer is redundant if it shares similarity with another neuron.[5]

Generally, removing a number of neurons from a fully connected layer results in compression of the data[6].But for the measure of degree of compression to be meaningful and to have direct connection with the remove of neurons, any unit with redundant functionality after training must be removed [2].That's when pruning based on distinctiveness can be useful.

Incorporating pruning technique ensures there is no duplication of neurons.Apart from saving the time and storage during the training process, pruning technique spares only the most significant neurons.Model consisting of only significant neurons is more sensitive toward further progressive reduction of the neurons leading to proportional compression.[7]

2 Method

As mentioned before, LeNet-5 Architecture published by LeCun et al., in 1998 is taken to be the baseline model and trained using preprocessed chosen part of vehicle-x dataset for classification.Preprocessing is done to the chosen part vehicle-x dataset as described in dataset subsection of introduction section.Finally, Testing accuracy of the baseline model is computed.

Further,to understand the effect of pruning in the performance of the trained model, pruning is implemented in the sixth layer which is a fully connected layer with 84 units. Pruning of an neuron in the fully connected layer is done one by one (followed by training the model again) and the performance of the model is assessed for every time a neuron is pruned.A neuron that shares maximum similarity with another neuron based on cosine similarity of their respective activation vectors is chosen for pruning at each step.

Testing accuracy is obtained every time a neuron is pruned followed by retraining of the model.These test accuracies are then plotted with respect to their corresponding number of neurons that are pruned.

Algorithm/(pseudocode) for pruning an neuron for this model is as mentioned below.

Initialize MaxCosineSimilarity to be 0
i $\Longrightarrow$ (0,number of classes)
j $\Longrightarrow$ (i, number of classes)
-
compute cosine similarity (i,j)
if absolute(cosine similarity (i,j)) > MaxCosineSimilarity
pruned = i
added = j
-
fc weight j $\Longrightarrow$ fc weight j + fc weight i
(fc weight i) is pruned
fc bias j $\Longrightarrow$ fc bias j + fc bias i
(fc bias i) is pruned
(output weight i) is pruned

3 Results

For the baseline model which is LeNet-5 Architecture , the testing accuracy achieved is 0.57 .Training and validation accuracy during the training process is plotted in the Fig. 3.

Testing Accuracies of pruned networks are plotted with respect to their number of neurons that are pruned in the Fig. 4 .It's evident that there is a significant fall in testing accuracy in the model when the number of neurons that are pruned is more than 50. This explains there are approximately 50 redundant neurons in the penultimate fully connected layer (the sixth layer) of the model.

The goal of this study has been achieved which is studying of effects of pruning in a fully connected layer of a CNN classification model.From the results,it's understood that as neurons are pruned from the fully connected layer one by one, there is no difference in testing accuracy of the model initially.But when it's number of neurons that are pruned exceeds is 50, model performs so badly which is evident from falling of testing accuracy of the model.

4 Conclusion and future work

From the experiment, it can be concluded that model whose fully connected layer is pruned, performs so badly after a threshold point where there are no more redundant neurons.Removal of neurons after that will results in significant loss in information due to which model's prediction ability deteriorates.

The future work pertaining to this experiment will be to incorporate advanced data analysis in order to implement more robust data preparation for better results.In this experiment, only one type of pruning which is based on distinctiveness has been explored.Exploring more complex pruning algorithm would be one of the key things to look forward to further take this experiment forward.

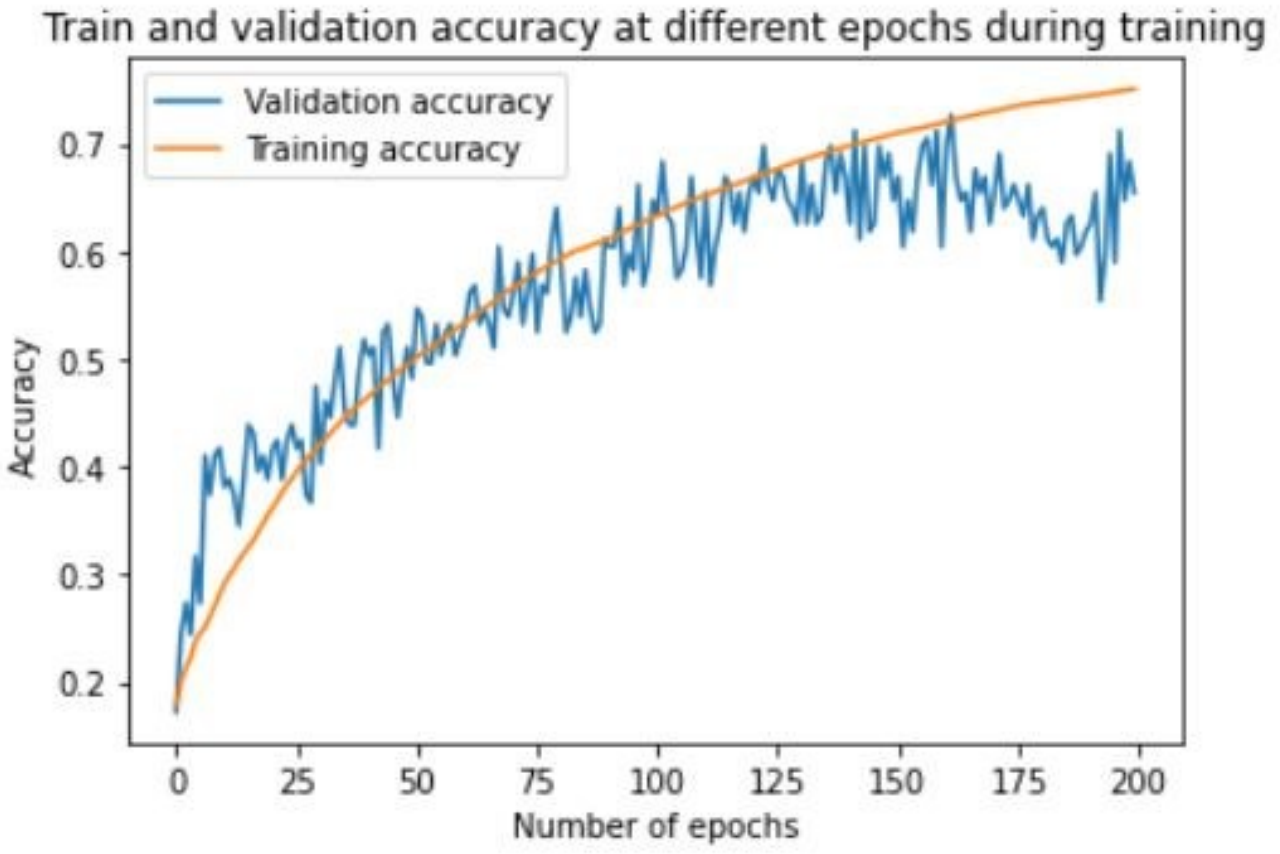

Fig. 3. Baseline architecture (LeNet-5) training-validation accuracy during training.

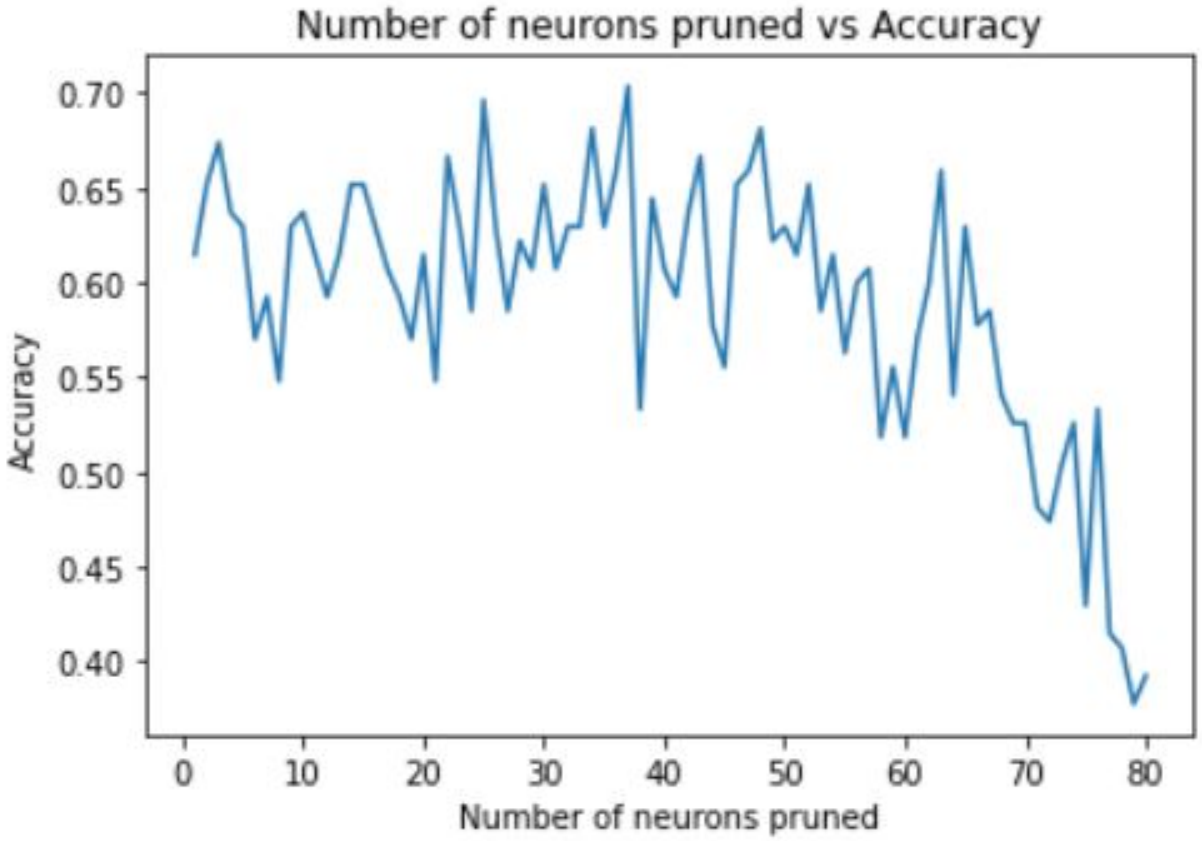

Fig. 4. Testing Accuracies of pruned networks vs number of neurons that are pruned

References

1.Zheng, L., Gedeon, T., Yao, Y., Yang, X. and Naphade, M.:Simulating Content Consistent Vehicle Datasets with Attribute Descent.(2020).
2.Gedeon, T. and Harris, D., n.d.: Progressive image compression. IJCNN International Joint Conference on Neural Networks.(1992).
3.LeNet-5-A Classic CNN Architecture, https://www.datasciencecentral.com/profiles/blogs/lenet-5-a-classic-cnn-architecture.
4.Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE. 86, 2278-2324 (1998).
5.SHAMIR, N., SAAD, D., MAROM, E.: NEURAL NET PRUNING BASED ON FUNCTIONAL BEHAVIOR OF NEURONS. International Journal of Neural Systems. 04, 143-158 (1993).
6.Karnin, E.: A simple procedure for pruning back-propagation trained neural networks. IEEE Transactions on Neural Networks. 1, 239-242 (1990).
7.Gedeon, T. and Harris, D.: Network Reduction Techniques. Int. Conf. on Neural Networks Methodologies and Applications, AMSE, San Diego. 2, 25-34 (1991).