Wide Residual Networks for CIFAR-10 Image Classification

Siri Pranitha Namburi
Department of Computer Science
Texas A & M University
College Station, TX 77801
siri@tamu.edu

Abstract

Residual networks(Resnet[1]) were shown to be able to scale up to thousand times and still have improving performance without the degradation problem. But as the accuracy increased, for further improvement, more than double the original layers were required which makes the residual networks slow to train. One proposed solution to this is the Wide Residual networks[3], an architecture where increasing the width and decreasing the depth of the residual network has shown better performance than the original ResNet[1]. My Project approaches at using the Wide Resnet architecture and uses Data Augumentation, learning rate scheduling and Dropout regularisation to successfully train a model that reaches 92.5 percent accuracy on the CIFAR 10[2] data set.

1 Introduction

Latest residual networks had a large success in field of image classification including winning Imagenet and COCO 2015 competition as well as achieving state of the art bench marks. Compared to deep neural networks they show better generalization and less degradation as well as speeding up convergence. But accuracy improvement in deep residual networks required a substantially large number of of layers. Wide residual network achieves better performance by widening up the convolution layers in the residual blocks. Authors of the wide residual networks[3] had experimentally shown that the WRN-40-4, a 40 layer deep and 4 times wider residual network outperforms ResNet-1001 , a 1001 layer deep residual network while having comparable number of parameters. It can also be noted that it significantly faster to train wide residual networks as compared to original thin residual networks.

2 Methods

The main components used for this project are wide residual networks, data augmentation, Stochastic gradient descent with cosine annealing learning rate scheduler and Dropout. These strategies helped in improvement of accuracy as compared to original residual networks

2.1 Wide residual networks

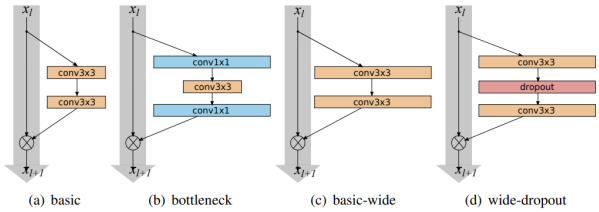

Figure 1: Figure 1: Wide Residual Networks from [4]. Fig(a) shows regular ResNets, Fig(c) shows the structure of basic-wide ResNets used in this Project

Training deeper residual network was leading to marginal improvement in accuracy. At the same time however, the residual block with the identity mapping that is allowing us to train very deep is at the same time a weakness of residual networks, as it is very possible that the gradient can flow through the network without learning anything. Hence a natural extension is widening the network. In particular, the representation in the residual blocks can be increased by adding more convolution layers per block, widening the convolution layers by adding more feature planes and increasing the filter sizes of the convolution layers. Due to skip connections , only some of the blocks in the regular residual networks will learn the essential features. Hence going deep would not cause much improvement in accuracy. On the other hand, going to a wider network results in a better representation of the input features and greater accuracy in a shallower network.[3]

2.2 Data augmentation

The Convolution Neural Network is not rotation invariant. Hence to make it more generalize, data augmentation techniques like random crop and random flip are done during the training phase. For random crop , the image of size 32x32 pixel with 3 channels is padded 4X4 padding and randomly cropped such that the final image is the same size as the original image. A horizontal flip with probability of 50 percent is applied on the image. Normalization is also done before feeding the image into the network

2.3 Gradient Descent

Stochastic gradient descent is used as an optimizer function and Cross entropy loss for the loss function. It is observed that adaptive learning rate is better for optimising the model. An initial learning rate of 0.01 is taken and a cosine annealing learning rate[5] is used to get good learning rate. The cosine learning rate scheduler helps in better and faster convergence of the model. A weight decay of $2\text{x}10^{-4}$ is used to handle the noise and get a more generalised model.

2.4 Dropout

Dropout is a technique for improving deep neural networks by reducing over fitting[4]. We switch of p fraction of the neurons randomly while training the neural network. It has an effect of regularisation on the network. In the current project p=0.5 is taken, i.e , 50 percent dropout is used on the final fully connected layer. From results we can see that the dropout increased the accuracy of the model on the public test data.

2.5 CIFAR -10 data

The model proposed in this project is tested on the CIFAR-10 dataset[2]. CIFAR-10 dataset has 50,000 images available for training and 10,000 images for testing purposes. The classes of the dataset with 10 random images of each is shown in figure2

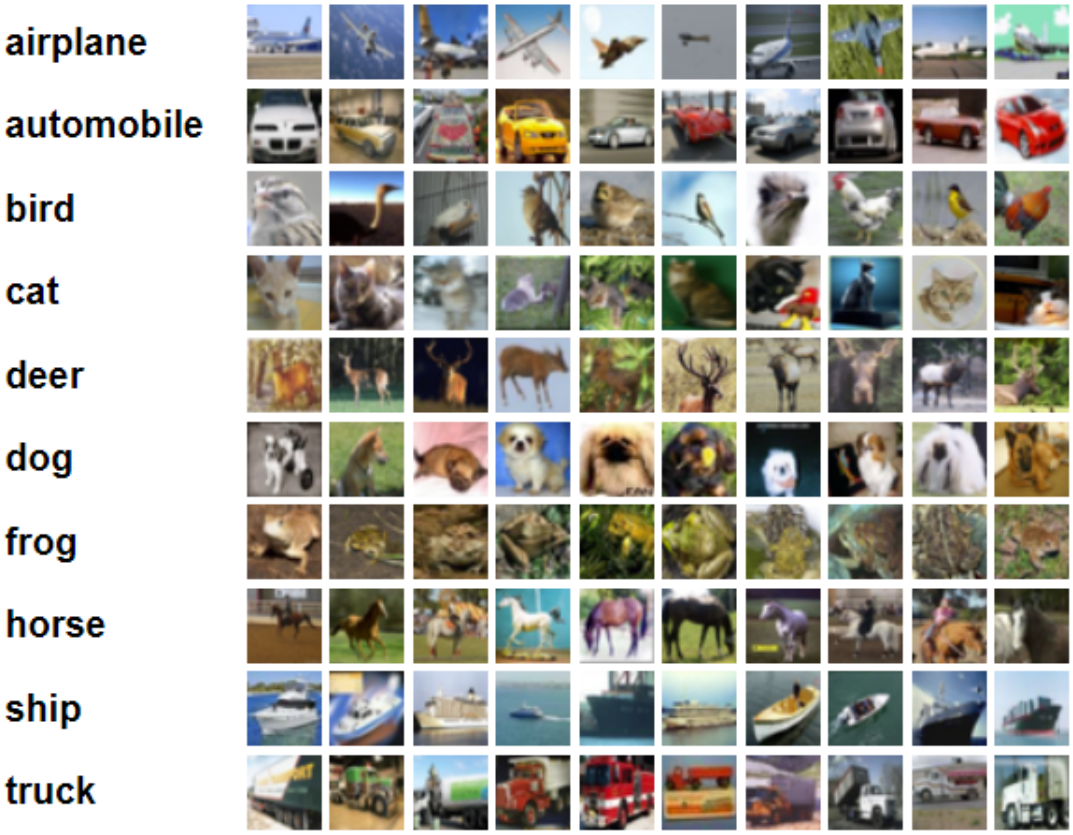

Figure 2: Figure 2: CIFAR-10 images with classes visualized[2]

3 Implementation

In this project , wide residual network is implemented where depth of the residual network is kept short(Residual network 18) and the width to be 2 and 3. Dropout rate is fixed to 50 percent in the training phase and a cosine rate scheduler is used. The architecture details of both the models are given below.

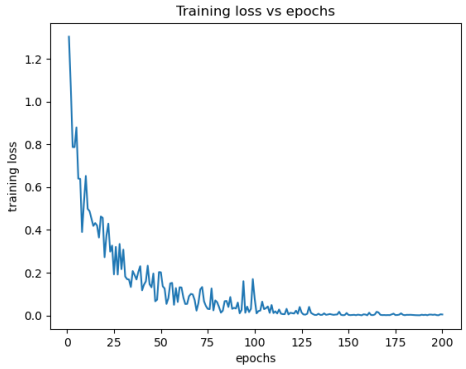

Figure 3: Figure 3: WRN-18-2 with width = 2, best model giving 91.97 percent accuracy. Trained for 200 epochs on training dataset

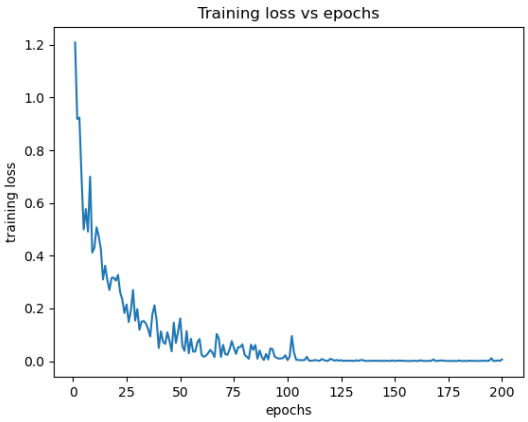

Figure 4: Figure 4: WRN-18-3 with width = 3, best model giving 92.5 percent accuracy. Trained for 200 epochs on training dataset

The training graph is shown in the Figure 3 which shows loss for each image for WRN-18-3. Figure 4 shows training loss for each epoch used for the final model WRN-18-3. It was trained for 200 epochs. I used Pytorch version 1.10.2 to train the model on Grace cluster. The model took approximately 2 hours to run 200 epochs

4 Results

The following table 1 shows the results on the public test dataset of CIFAR-10 data.

Network	Details	Test accuracy
ResNet-18	Data Augumentation+Cosine Learning Rate	89.84
WRN-18-2	Data Augumentation+Cosine Learning Rate	91.63
WRN-18-3	Data Augumentation+Cosine Learning Rate	91.83
WRN-18-2	Data Augumentation+Cosine Learning Rate+Dropout	91.97
WRN-18-3	Data Augumentation+Cosine Learning Rate+Dropout	92.50

Table 1: Result

5 Conclusion and Future work

In this project, Wide Residual Network is implemented and we are able to achieve 92.5 percent accuracy on public test of CIFAR-10 data. We are able to show that learning rate scheduling, dropout and data augumentation strategies along with wide residual network led to considerable improvement in accuracy as compared to the original residual network. In future work, we can make use of more preprocessing strategies like rotation of input image, adding gaussian noise and so on to make the data augumentation more robust and improve the accuracy of the model

6 References

[1] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[2] Krizhevsky, Alex, and Geoffrey Hinton. "Learning multiple layers of features from tiny images." (2009): 7.

[3] Zagoruyko, Sergey, and Nikos Komodakis. "Wide residual networks." arXiv preprint arXiv:1605.07146 (2016).

[4] Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." The journal of machine learning research 15.1 (2014): 1929-1958.

[5] Loshchilov, Ilya, and Frank Hutter. "Sgdr: Stochastic gradient descent with warm restarts." arXiv preprint arXiv:1608.03983 (2016).