Fashion Mnist Project Report

Tianjin International Engineering Institute, Tianjin University, Nankai, Tianjin, China, 300072
houys@tju.edu.cn

Abstract

The method of residual learning has solved the problem of neural network degradation and achieved great success in the field of deep learning.This paper is based on the 18-layer residual network, adopts the image augmentation strategy of Random Erasing and Random Image Cropping and Patching, then use the SGD optimizer to train the model, and the accuracy of the final model reaches 0.9588.

1 Introduction

The deep network structure contains all possible solution Spaces of the shallow network structure, but in the actual network training, with the increase of the network depth, the network accuracy appears saturation or even decline.He et al. proposed a deep residual network to solve this problem. [2].This network makes it possible to train deeper network structures.

As important as model structure is data augmentation technology,data augmentation can expand samples, prevent overfitting, and improve the robustness of models. Common methods include Random flipping, DropOut, Random Cropping, Batch Normallization, etc. However, these traditional data augmentation methods bring limited model gains, so Zhong et al. proposed the method of Random Erasing [10], and Takahashi et al. proposed the method of Random Image Cropping and Patching [8].These two methods inject new vitality into the image augmentation strategy.

BatchNorm can make the input of each layer of neural network keep the same distribution during deep neural network training, but the Batch size will affect the performance of the normalization, it usually performs better on large Batch sizes.Thus, Saurabh Singh et al. proposed another Normalization method, Filter Response Normalization Layer [6], which could remove the dependency on Batch size.

Optimization algorithm is also very important, and there are various, among which SGD and Adam are representative.SGD is often used in late tuning, but the problem with SGD is slow convergence in the early stage.In the late stage of Adam's adaptive learning rate training, the learning rate appears to be extreme. When updating parameters, the learning rate in some dimensions is particularly large, while the learning rate in some dimensions is particularly small.Luo et al. recently proposed an optimizer Adabound [4] that can be as fast as Adam and as good as SGD. In addition, the work [5] of Sebastian Ruder et al and the work [7] of Ruoyu Sun et al sorted out the development of optimization algorithms and compared each optimizer theoretically and experimentally.

The work [3] of He et al. summarized some of the tricks for training neural network models, which are very helpful. The work [1] of Dong et al. and the work [9] of Yamada et al. are instructive for training a deep residual network.

2 Model Configuration

2.1 Architecture

The model architecture is shown in the figure 1

Stages shown in figure 2

2.2 Configuration

Figure size: 28

Batch size: 256

Optimizer: $SGD(momentum = 0.9, weigh\_decay = 5e - 4)$

Initial learning rate: $lr = 0.1 * batch_size/256$

Loss function: CrossEntropyLoss

Data augmentation: Random Erasing + RICAP(beta=0.1)

3 Experiment Framework

3.1 Implementation Details

Learning rate warmup: first, a small Learning rate training of 0.01 was conducted to train 10 epochs. During the training, the Learning rate was gradually increased to the initial Learning rate.

Train: CosineAnnealingLR scheduler was used to Train 100 epochs;

Warm retrain: using CosineAnnealingWarmRestarts scheduler, with a smaller initial learning rate 0.001, retrain the model 10 epoch to make a small adjustment.

In the whole training, Early stopping is set to 15.

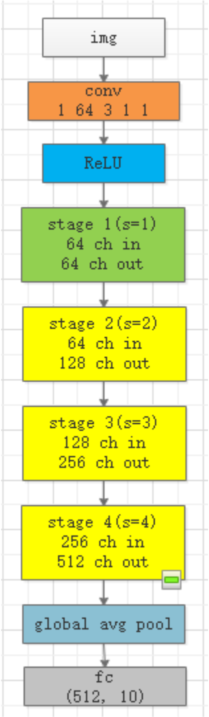
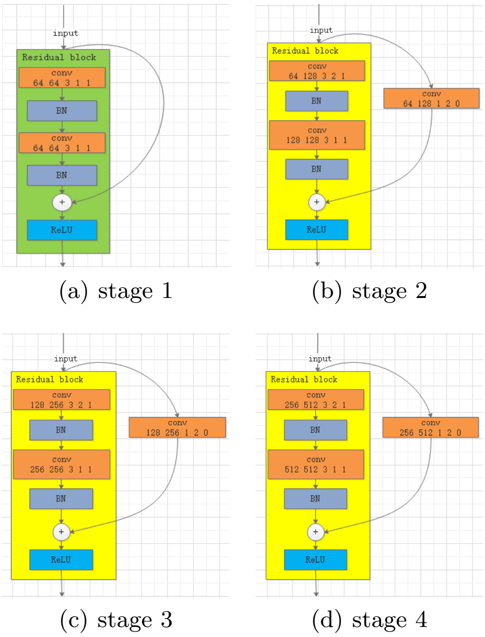


Figure 1: The model arrchitecture



(a) stage 1

(b) stage 2

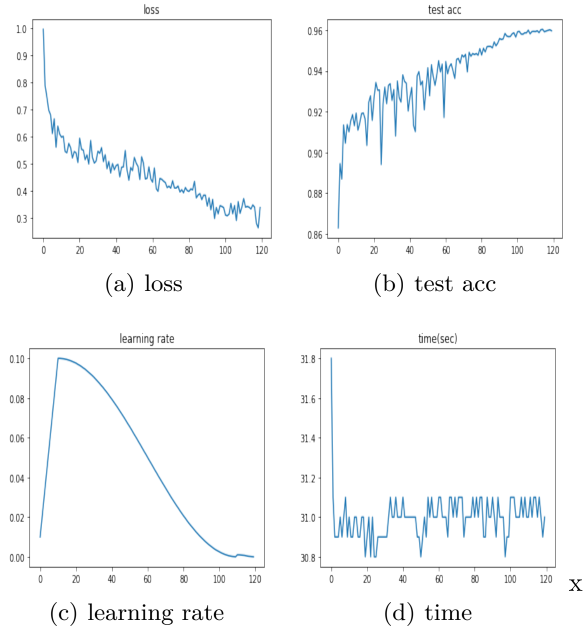
(c) stage 3

(d) stage 4

Figure 2: Four stages

3.2 Evaluation

Evaluation curves can be seen in figure ??



(a) loss

(b) test acc

(c) learning rate

(d) time

Figure 3: Evaluations

3.3 Discussion

During the training process, there is a problem that I haven't figured out yet, that is, the accuracy of training is not so high. The figure 4 is my training accuracy curve.But the accuracy of the test is relatively high.

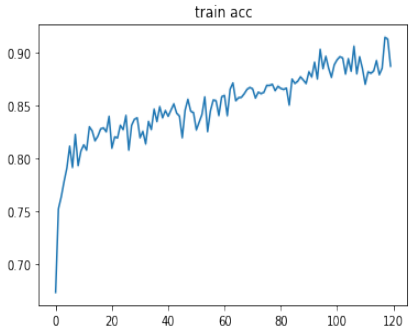


Figure 4: train acc curve

4 Conclusion

In our work, we adopt the classic deep residual network and adopt the data augmentation strategy of Random Erasing and Random Image Cropping and Patching. Experimental results show that this strategy performs well

Acknowledgments

Baseline: $https : //github.com/monkeyDemon/Learn_Dive-into-DL-PyTorch$

References

[1] Xuanyi Dong, Guoliang Kang, Kun Zhan, and Yi Yang. Eraserelu: A simple way to ease the training of deep convolution neural networks. CoRR, abs/1709.07634, 2017.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. CoRR, abs/1512.03385, 2015.

[3] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. CoRR, abs/1812.01187, 2018.

[4] Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. Adaptive gradient methods with dynamic bound of learning rate. CoRR, abs/1902.09843, 2019.

[5] Sebastian Ruder. An overview of gradient descent optimization algorithms. CoRR, abs/1609.04747, 2016.

[6] Saurabh Singh and Shankar Krishnan. Filter response normalization layer: Eliminating batch dependence in the training of deep neural networks, 2019.

[7] Ruoyu Sun. Optimization for deep learning: theory and algorithms, 2019.

[8] Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara. Data augmentation using random image cropping and patching for deep cnns. CoRR, abs/1811.09030, 2018.

[9] Yoshihiro Yamada, Masakazu Iwamura, and Koichi Kise. Shakedrop regularization. CoRR, abs/1802.02375, 2018.

[10] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. CoRR, abs/1708.04896, 2017.