Graph-based Knowledge Distillation by Multi-head Attention Network :Supplementary Material

Seunghyun Lee
lsh910703@gmail.com
Byung Cheol Song
bcsong@inha.ac.kr
Inha University
Incheon, Republic of Korea

1 Network Architecture

This section describes the network architectures used in this paper. We adopted VGG [7], WResNet [8], ResNet [3], and MobileNet [4] as shown in Fig. 1. We sensed feature maps at the front and back of the dotted box, and used the sensed results as input to the multi-head graph distillation (MHGD) module.

When the experimental result for TinyImageNet [2] is obtained (Table 2 in the main paper), max pooling was added after the fourth convolutional layer block in the VGG architecture. In the WResNet architecture, the stride of the first convolutional layer was set to 2.

In addition, we use modified VGG network which have the feature map of the same size as WResNet-Teacher for obtaining Table 3.

2 Training Setting

All algorithms were implemented using Tensorflow [1]. Also, weights of all networks were initialized with He's initialization [3] and L_2 regularization was applied. A stochastic gradient descent (SGD) [5] was used as the optimizer and a Nesterov accelerated gradient [6] was applied. All numerical values in the tables and figures are the averages of the total five trials.

Next, we explain the augmentation of the dataset. All datasets are normalized to have a range of [-0.5, 0.5], and horizontal random flip is used for augmentation. Also, the images of CIFAR100 are zero-padded by 4 pixels, and the images of Tiny-ImageNet are zero-padded by 8 pixels. Then the zero-padded images are randomly cropped to the original size.

Next, we describe the hyper-parameters we used for network learning. First, the hyper-parameters used in the learning of CIFAR100 and TinyImageNet to obtain the experimental results of Table 1 and 2 of this paper are as follows. In case of VGG, learning was proceeded for 200 epochs and an initial learning rate was set to 0.01, which is reduced by 0.1 times at 100 and 150 epochs. In WResNet, learning was proceeded for 200 epochs and an initial learning rate was set to 0.1, which is reduced by 0.2 times at 60, 120 and 160 epochs, respectively. Because WResNet converges relatively quickly, we halved the training epoch of

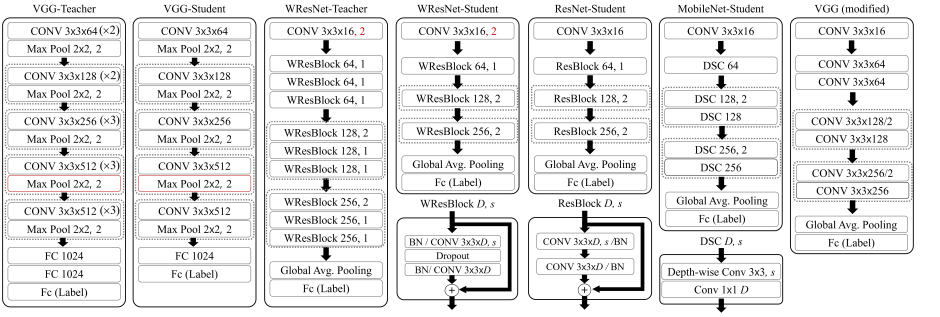

Figure 1: The block diagram for network architectures used in the proposed scheme.

the student network. The batch size of all networks was set to 128, and the weight decay of L_2 regularization was fixed to 5×10^{-4}.

In Table 3, the hyper-parameters of MobileNet and ResNet were the same as those of WResNet. In Table 4, we used the same VGG network and hyper-parameters as those used in Table 1, and only changed the number of attention heads A for the ablation study. The following describes hyper-parameters for learning of the multi-head attention network (MHAN). Basically, we use the same hyper-parameters as when learning CNN. However, the learning rate was fixed at 0.1, and only 20 epochs were learned. In all cases except for the ablation study, the number of attention heads of the networks was 8.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[5] Jack Kiefer, Jacob Wolfowitz, et al. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

[6] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In *Doklady AN USSR*, volume 269, pages 543–547, 1983.

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[8] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.