Densely Guided Knowledge Distillation using Multiple Teacher Assistants Supplementary Material

Wonchul Son[1], Jaemin Na[1], Junyong Choi[1], Wonjun Hwang[1]
[1]Ajou University, Republic of Korea
{dnjscjf92, osial46, chldusxkr, wjhwang}@ajou.ac.kr

1. Network Architecture Details

In this section, we explain the detail network architectures of the teacher, the teacher assistants, and the student models used in our experiments such as Plain CNN [2], ResNet [1], WRN [4] and VGG [3] using CIFAR-10, CIFAR-100 and ImageNet datasets.

1.1. Plain CNN model configurations

We used the Plain CNN network structures as same as used in the Teacher Assistant Knowledge Distillation (TAKD) [2] to compare with our proposed method, Densely Guided Knowledge Distillation (DGKD). Under Table 1 shows teacher (T_{10}), teacher assistants (A_8, A_6, A_4) and student (S_2) structures used in experiments Table 1, Table 4, and Table 6 in this paper. Additionally, for the Table 5 experiment in our paper, the other assistants (A_9, A_7, A_5, A_3) are made from networks (T_{10}, A_8, A_6, A_4) by excluding the last convolutional layer respectively.

Table 1. Plain CNN model configurations for CIFAR-100.

T_{10}	A_8	A_6	A_4	S_2
$\begin{bmatrix} Conv3\times3-32 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-32 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-32 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-32 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-32 \\ BN, ReLU \end{bmatrix} \times 1$
maxpool				
$\begin{bmatrix} Conv3\times3-64 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-64 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-64 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-64 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-32 \\ BN, ReLU \end{bmatrix} \times 1$
maxpool				
$\begin{bmatrix} Conv3\times3-128 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-128 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-128 \\ BN, ReLU \end{bmatrix} \times 2$	-	-
maxpool				
$\begin{bmatrix} Conv3\times3-256 \\ BN, ReLU \end{bmatrix} \times 4$	$\begin{bmatrix} Conv3\times3-256 \\ BN, ReLU \end{bmatrix} \times 2$	-		
maxpool				
fc-512	fc-64			
fc-100				
softmax				

1.2. ResNet model configurations for CIFAR

We used the standard ResNet [1] structure for experiments Table 2, Table 6, Table 7 and Table 8 in our paper. Additionally, we selected the teacher, the teacher assistant, and the student to have same layer's depth gap sizes for the multi-step distillation paths.

Table 2. ResNet model configurations for CIFAR-10.

T_{26}	A_{20}	A_{14}	S_8
$\begin{bmatrix} Conv3\times3-16 \\ BN, ReLU \end{bmatrix}$			
$\begin{bmatrix} Conv3\times3-16 \\ BN, ReLU \\ Conv3\times3-16 \\ BN, ReLU \end{bmatrix} \times 4$	$\begin{bmatrix} Conv3\times3-16 \\ BN, ReLU \\ Conv3\times3-16 \\ BN, ReLU \end{bmatrix} \times 3$	$\begin{bmatrix} Conv3\times3-16 \\ BN, ReLU \\ Conv3\times3-16 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-16 \\ BN, ReLU \\ Conv3\times3-16 \\ BN, ReLU \end{bmatrix} \times 1$
$\begin{bmatrix} Conv3\times3-32 \\ BN, ReLU \\ Conv3\times3-32 \\ BN, ReLU \end{bmatrix} \times 4$	$\begin{bmatrix} Conv3\times3-32 \\ BN, ReLU \\ Conv3\times3-32 \\ BN, ReLU \end{bmatrix} \times 3$	$\begin{bmatrix} Conv3\times3-32 \\ BN, ReLU \\ Conv3\times3-32 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-32 \\ BN, ReLU \\ Conv3\times3-32 \\ BN, ReLU \end{bmatrix} \times 1$
$\begin{bmatrix} Conv3\times3-64 \\ BN, ReLU \\ Conv3\times3-64 \\ BN, ReLU \end{bmatrix} \times 4$	$\begin{bmatrix} Conv3\times3-64 \\ BN, ReLU \\ Conv3\times3-64 \\ BN, ReLU \end{bmatrix} \times 3$	$\begin{bmatrix} Conv3\times3-64 \\ BN, ReLU \\ Conv3\times3-64 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-64 \\ BN, ReLU \\ Conv3\times3-64 \\ BN, ReLU \end{bmatrix} \times 1$
avgpool			
fc-10			
softmax			

Table 3. ResNet model configurations for CIFAR-100.

T_{56}	A_{44}	A_{32}	S_{20}
$\begin{bmatrix} Conv3\times3-16 \\ BN, ReLU \end{bmatrix}$			
$\begin{bmatrix} Conv3\times3-16 \\ BN, ReLU \\ Conv3\times3-16 \\ BN, ReLU \end{bmatrix} \times 9$	$\begin{bmatrix} Conv3\times3-16 \\ BN, ReLU \\ Conv3\times3-16 \\ BN, ReLU \end{bmatrix} \times 7$	$\begin{bmatrix} Conv3\times3-16 \\ BN, ReLU \\ Conv3\times3-16 \\ BN, ReLU \end{bmatrix} \times 5$	$\begin{bmatrix} Conv3\times3-16 \\ BN, ReLU \\ Conv3\times3-16 \\ BN, ReLU \end{bmatrix} \times 3$
$\begin{bmatrix} Conv3\times3-32 \\ BN, ReLU \\ Conv3\times3-32 \\ BN, ReLU \end{bmatrix} \times 9$	$\begin{bmatrix} Conv3\times3-32 \\ BN, ReLU \\ Conv3\times3-32 \\ BN, ReLU \end{bmatrix} \times 7$	$\begin{bmatrix} Conv3\times3-32 \\ BN, ReLU \\ Conv3\times3-32 \\ BN, ReLU \end{bmatrix} \times 5$	$\begin{bmatrix} Conv3\times3-32 \\ BN, ReLU \\ Conv3\times3-32 \\ BN, ReLU \end{bmatrix} \times 3$
$\begin{bmatrix} Conv3\times3-64 \\ BN, ReLU \\ Conv3\times3-64 \\ BN, ReLU \end{bmatrix} \times 9$	$\begin{bmatrix} Conv3\times3-64 \\ BN, ReLU \\ Conv3\times3-64 \\ BN, ReLU \end{bmatrix} \times 7$	$\begin{bmatrix} Conv3\times3-64 \\ BN, ReLU \\ Conv3\times3-64 \\ BN, ReLU \end{bmatrix} \times 5$	$\begin{bmatrix} Conv3\times3-64 \\ BN, ReLU \\ Conv3\times3-64 \\ BN, ReLU \end{bmatrix} \times 3$
avgpool			
fc-100			
softmax			

1.3. ResNet model configurations for ImageNet

For the ImageNet experiments on Table 3 in our paper, we designed the ResNet 26-layer (A_{26}) similar with the structure of ResNet 18-layer (S_{18}), considering the same layer's depth gap sizes for the multi-step distillation path.

Table 4. ResNet model configurations for ImageNet.

T_{34}	A_{26}	S_{18}
$\begin{bmatrix} Conv7\times7-64 \\ BN, ReLU \end{bmatrix}$		
maxpool		
$\begin{bmatrix} Conv3\times3-64 \\ BN, ReLU \\ Conv3\times3-64 \\ BN, ReLU \end{bmatrix} \times 3$	$\begin{bmatrix} Conv3\times3-64 \\ BN, ReLU \\ Conv3\times3-64 \\ BN, ReLU \end{bmatrix} \times 3$	$\begin{bmatrix} Conv3\times3-64 \\ BN, ReLU \\ Conv3\times3-64 \\ BN, ReLU \end{bmatrix} \times 2$
$\begin{bmatrix} Conv3\times3-128 \\ BN, ReLU \\ Conv3\times3-128 \\ BN, ReLU \end{bmatrix} \times 4$	$\begin{bmatrix} Conv3\times3-128 \\ BN, ReLU \\ Conv3\times3-128 \\ BN, ReLU \end{bmatrix} \times 3$	$\begin{bmatrix} Conv3\times3-128 \\ BN, ReLU \\ Conv3\times3-128 \\ BN, ReLU \end{bmatrix} \times 2$
$\begin{bmatrix} Conv3\times3-256 \\ BN, ReLU \\ Conv3\times3-256 \\ BN, ReLU \end{bmatrix} \times 6$	$\begin{bmatrix} Conv3\times3-256 \\ BN, ReLU \\ Conv3\times3-256 \\ BN, ReLU \end{bmatrix} \times 3$	$\begin{bmatrix} Conv3\times3-256 \\ BN, ReLU \\ Conv3\times3-256 \\ BN, ReLU \end{bmatrix} \times 2$
$\begin{bmatrix} Conv3\times3-512 \\ BN, ReLU \\ Conv3\times3-512 \\ BN, ReLU \end{bmatrix} \times 3$	$\begin{bmatrix} Conv3\times3-512 \\ BN, ReLU \\ Conv3\times3-512 \\ BN, ReLU \end{bmatrix} \times 3$	$\begin{bmatrix} Conv3\times3-512 \\ BN, ReLU \\ Conv3\times3-512 \\ BN, ReLU \end{bmatrix} \times 2$
avgpool		
fc-1000		
softmax		

1.4. Wide Residual Network model configurations for CIFAR

We used the standard Wide Residual Network [4] (WRN) for experiments on table 8 in our paper and fixed the widen factor as 2. We used WRN40×2 as teacher, WRN34×2, WRN28×2, WRN22×2 as teacher assistants and WRN16×2 as student model considering the same layer's depth gap sizes for the multi-step distillation path.

Table 5. WRN model configurations for CIFAR-100; for simplicity, we omitted the three convolutional skip connections.

$T_{40\times2}$	$A_{34\times2}$	$A_{28\times2}$	$A_{22\times2}$	$S_{16\times2}$
$\begin{bmatrix} BN, ReLU \\ Conv3\times3-16 \end{bmatrix}$				
$\begin{bmatrix} BN, ReLU \\ Conv3\times3-32 \\ BN, ReLU \\ Conv3\times3-32 \end{bmatrix} \times 6$	$\begin{bmatrix} BN, ReLU \\ Conv3\times3-32 \\ BN, ReLU \\ Conv3\times3-32 \end{bmatrix} \times 5$	$\begin{bmatrix} BN, ReLU \\ Conv3\times3-32 \\ BN, ReLU \\ Conv3\times3-32 \end{bmatrix} \times 4$	$\begin{bmatrix} BN, ReLU \\ Conv3\times3-32 \\ BN, ReLU \\ Conv3\times3-32 \end{bmatrix} \times 3$	$\begin{bmatrix} BN, ReLU \\ Conv3\times3-32 \\ BN, ReLU \\ Conv3\times3-32 \end{bmatrix} \times 2$
$\begin{bmatrix} BN, ReLU \\ Conv3\times3-64 \\ BN, ReLU \\ Conv3\times3-64 \end{bmatrix} \times 6$	$\begin{bmatrix} BN, ReLU \\ Conv3\times3-64 \\ BN, ReLU \\ Conv3\times3-64 \end{bmatrix} \times 5$	$\begin{bmatrix} BN, ReLU \\ Conv3\times3-64 \\ BN, ReLU \\ Conv3\times3-64 \end{bmatrix} \times 4$	$\begin{bmatrix} BN, ReLU \\ Conv3\times3-64 \\ BN, ReLU \\ Conv3\times3-64 \end{bmatrix} \times 3$	$\begin{bmatrix} BN, ReLU \\ Conv3\times3-64 \\ BN, ReLU \\ Conv3\times3-64 \end{bmatrix} \times 2$
$\begin{bmatrix} BN, ReLU \\ Conv3\times3-128 \\ BN, ReLU \\ Conv3\times3-128 \end{bmatrix} \times 6$	$\begin{bmatrix} BN, ReLU \\ Conv3\times3-128 \\ BN, ReLU \\ Conv3\times3-128 \end{bmatrix} \times 5$	$\begin{bmatrix} BN, ReLU \\ Conv3\times3-128 \\ BN, ReLU \\ Conv3\times3-128 \end{bmatrix} \times 4$	$\begin{bmatrix} BN, ReLU \\ Conv3\times3-128 \\ BN, ReLU \\ Conv3\times3-128 \end{bmatrix} \times 3$	$\begin{bmatrix} BN, ReLU \\ Conv3\times3-128 \\ BN, ReLU \\ Conv3\times3-128 \end{bmatrix} \times 2$
BN,ReLU				
avgpool				
fc-100				
softmax				

1.5. VGG model configurations for CIFAR

We used the standard VGG [3] network for experiments on table 8 in our paper, we used VGG13 (T_{13}), VGG11 (A_{11}), VGG8 (S_8) which is commonly used in other knowledge distillation papers.

Table 6. VGG model configurations for CIFAR-100.

T_{13}	A_{11}	S_8
$\begin{bmatrix} Conv3\times3-64 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-64 \\ BN, ReLU \end{bmatrix} \times 1$	$\begin{bmatrix} Conv3\times3-64 \\ BN, ReLU \end{bmatrix} \times 1$
maxpool		
$\begin{bmatrix} Conv3\times3-128 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-128 \\ BN, ReLU \end{bmatrix} \times 1$	$\begin{bmatrix} Conv3\times3-128 \\ BN, ReLU \end{bmatrix} \times 1$
maxpool		
$\begin{bmatrix} Conv3\times3-256 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-256 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-256 \\ BN, ReLU \end{bmatrix} \times 1$
maxpool		
$\begin{bmatrix} Conv3\times3-512 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-512 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-512 \\ BN, ReLU \end{bmatrix} \times 1$
maxpool		
$\begin{bmatrix} Conv3\times3-512 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-512 \\ BN, ReLU \end{bmatrix} \times 2$	$\begin{bmatrix} Conv3\times3-512 \\ BN, ReLU \end{bmatrix} \times 1$
maxpool		
fc-512		
fc-512		
fc-100		
softmax		

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 770–778, Jun. 2016. 1, 2

[2] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh. Improved knowledge distillation via teacher assistant. *34th AAAI Conf. on Artificial Intelligence*, pages 5191–5198, Feb. 2020. 1

[3] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 4

[4] S. Zagoruyko and N. Komodakis. Wide residual networks. *British Machine Vision Conference*, pages 87.1–87.12, Sept. 2016. 1, 3