Supplementary Material for Local Correlation Consistency for Knowledge Distillation

Xiaojie Li[1][0000−0001−6449−2727], Jianlong Wu[2(✉)][0000−0003−0247−5221], Hongyu Fang[3][0000−0002−9945−9385], Yue Liao[4][0000−0002−2671−0655], Fei Wang[1][0000−0002−1024−5867], and Chen Qian[1][0000−0002−8761−5563]

[1]SenseTime Research
[2]School of Computer Science and Technology, Shandong University
[3]School of Electronics Engineering and Computer Science, Peking University
[4]School of Computer Science and Engineering, Beihang University
lixiaojie@sensetime.com, jlwu1992@sdu.edu.cn, fanghongyu@pku.edu.cn, liaoyue.ai@gmail.com, {wangfei,qianchen}@sensetime.com

We provide more quantitative and qualitative evaluations of our method in the following sections. First, we compare the proposed LKD with other methods on the CIFAR10 dataset. Secondly, we show more visualization examples of the attention mask to show the ability of our class-aware attention module to capture the class-specific foreground regions of the global feature maps. Then we visualize and compare the feature distributions of teacher network and the student networks supervised by the global-based and our local-based consistency losses to show the superiority of our methods to keep the consistency between teacher and student.

1 Evaluation on CIFAR10

CIFAR10 is a commonly adopted dataset for comparison, which contains $50,000$ training images and $10,000$ testing images with 32×32 resolution. The dataset has 10 classes, where each class has an equal number of images. Following the setting on CIFAR100 (Section 4.1 of the main paper), we use the same student-teacher network pairs and the training strategies. Table 1 summarizes the results. We can see that our LKD method achieves better results than other state-of-the-art methods under four different student-teacher network pairs, which demonstrates the effectiveness and robustness of our proposed local correlation consistency supervision and the class-aware attention module.

We also observe that, under the student-teacher pair of WRN-16-2 and WRN-40-2, they achieve comparable classification accuracy. Specifically, the student models trained by AT [3], CC [2] and our LKD achieve better results than the teacher model. The Top-1 accuracy of LKD is 94.81%, which surpasses the teacher by 0.28%. As has been confirmed in [1] that a weaker teacher is able to train a stronger student, our results are not surprising because the CE baseline of the student has similar accuracy to the teacher.

Table 1. Comparison of classification accuracy on CIFAR10. The best results of the student network are highlighted in bold

Teacher Net.	Student Net.	CE	KD	AT	SP	CC	LKD	Teacher
ResNet110	ResNet14	90.99	92.45	92.68	92.58	92.34	**92.96**	93.88
ResNet110	ResNet20	92.39	93.32	93.53	92.51	93.39	**93.65**	93.88
WRN-40-2	WRN-16-1	91.52	92.48	92.55	92.54	92.23	**92.59**	94.53
WRN-40-2	WRN-16-2	93.44	94.47	94.70	94.45	94.54	**94.81**	94.53

2 Visualization of Attention Masks

To further show how the CAAT module works, we randomly select more images from the training set of ImageNet, and collect the original feature maps and the corresponding attention masks generated by the trained CAAT module. Both the feature maps and the attention masks are selected from the third stage of the teacher network. Results are shown in Fig. 1. We can find the similar observation with the main paper. The informative regions, such as the heads of the objects, are assigned relatively high value. The confusing background regions and the class-irrelevant foreground regions, which has less contribution to the classification task, are assigned relatively low value. Applying the attention masks to the feature maps can help the student network to focus on those class-relevant regions and ignore these confusing regions in images.

3 Visualization of the Consistency between Teacher and Student.

To investigate the ability of our method to keep the correlation knowledge consistent between teacher and student, we visualize the feature distributions. We allocate the features from the middle layers of the teacher and the student networks. Then we compute the pairwise $\cos(\theta)$ similarities within the global features and the local features, respectively, after which we represent them using histogram distributions.

In our implementation, the features are gathered from the second block of ResNet110 and ResNet20 with a resolution of 16×16. We compare the distribution of four models trained on CIFAR100, including: (1) the teacher network trained by CE (ResNet110); (2) the student network trained by CE (ResNet20-CE) without the supervised of the teacher; (3) the student network trained by the proposed LKD proposed (ResNet20-LKD); (4) the student network trained by the global correlation based loss (ResNet20-GKD) by setting the $k = 1$ with our method. For each model, we can extract the global features and the local features, and compute 1×10^6 pairwise similarities. Then we visualize the global distribution and local distribution separately.

In Figure 2, we show and compare the distributions of the global and local features of these four models. For the distributions of student models, the more

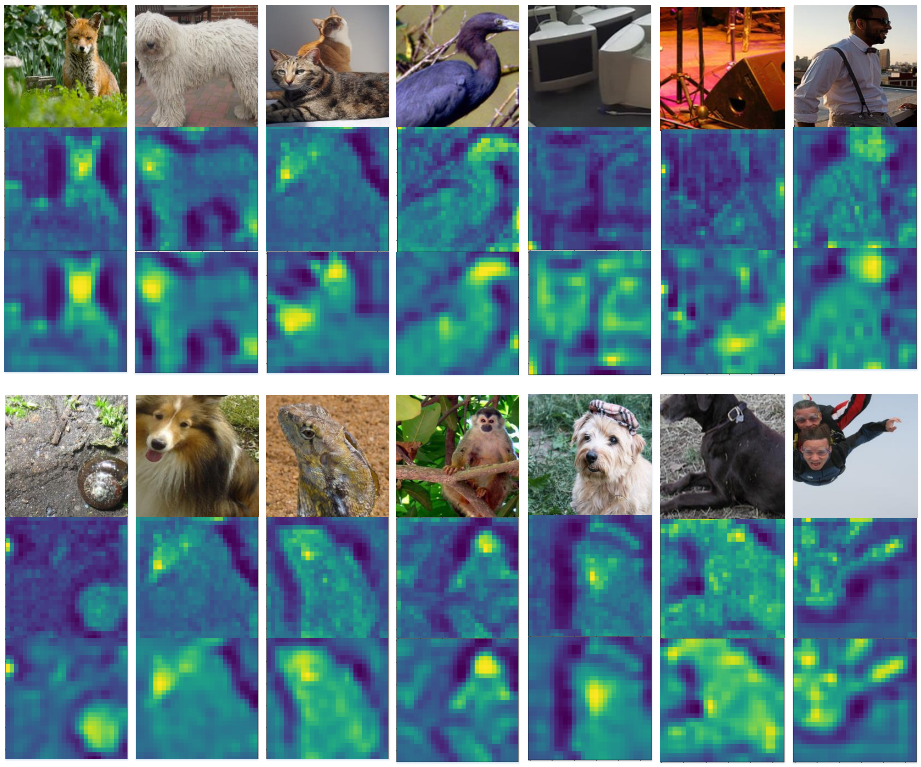

Fig. 1. Visualization of the attention maps. For each image, the original image is placed on the first row, the feature maps generated by the teacher model is placed on the second row, the corresponding attention masks generated by CAAT module is placed on the third row. High value is shown in yellow and low value in blue

overlaps with the teachers' distribution correspond to the better performance. We have the following observations. (1) According to the first column, there is a large gap between distributions of the teacher and the CE based student. (2) Results in the second column compare the global correlation based student with the teacher. It achieves better results but there still exists obvious difference, which can be attributed to the unsatisfying ability of global feature based relationships to transfer correlation information of the local regions. (3) The last column demonstrates the advantage of our proposed method, where the distributions are almost coincident. This can further verify our contribution that local feature based correlations can contribute more to transfer sufficient knowledge from the teacher network.

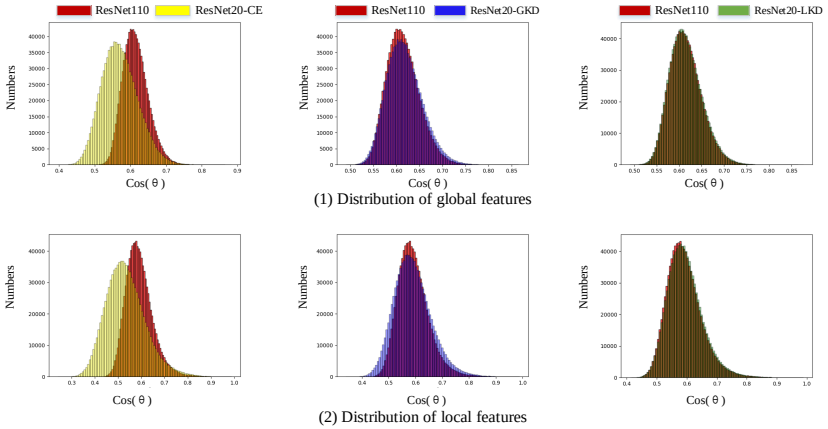

Fig. 2. cos(θ) histogram distributions of the global features (top) and the local features (bottom). $Hist_{Red}$ represents the teacher network, which is compared with each of the student networks from left to right column. $Hist_{Yellow}$ represents the student network trained by CE. $Hist_{Blue}$ represents the student network supervised by global feature based correlation

References

1. Furlanello, T., Lipton, Z.C., Tschannen, M., Itti, L., Anandkumar, A.: Born again neural networks. arXiv preprint arXiv:1805.04770 (2018)
2. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: IEEE ICCV. pp. 5007–5016 (2019)
3. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)