
SUPPLEMENTARY MATERIALS FOR BLACK-BOX KNOWLEDGE DISTILLATION

Anonymous authors

Paper under double-blind review

1 IMPLEMENTATION DETAILS

For CIFAR-100, we set the batch size as 64 and the base learning rate as 0.05. For ImageNet, we set the batch size as 128, and the base learning rate as 0.1. For MS-COCO, we set the batch size as 8 and the base learning rate as 0.01. We take 1 GPU to train the model on CIFAR-100 and 8 GPUs on ImageNet and MS-COCO. For prediction augmentation, we take $K = 5$ temperatures $[2.0, 3.0, 4.0, 5.0, 6.0]$ and denote the median of them as T ($T = 4.0$ here). We run each experiment five times and report the average results.

2 DATASETS

We take three widely researched datasets, 1) CIFAR-100 Krizhevsky et al. (2009), with 60,000 images in total (50,000 for training and 10,000 for validation) from 100 categories, the resolution of images is 32×32 , 2) ImageNet Russakovsky et al. (2015), one of the most important benchmark datasets for image classification, with nearly 1.3 million training images and 50,000 images for validation. The images come from 1,000 categories and are in high resolution, and 3) MS-COCO Lin et al. (2014), a mainstream dataset for object detection, with 118,000 training images and 5,000 validation images from 80 categories. CIFAR-100 and ImageNet are taken for image classification, and MS-COCO is utilized for object detection.

3 MORE ANALYSES

Comparison with Teacher Model It is interesting to explore the performance gap between the student and teacher model in our method. In Table 1, we calculate the gap between the accuracies of the student and teacher model, we note that the gap is negative when the student model outperforms the teacher model. We can observe that with our carefully designed method, the student model performance is highly close to the teacher model, with an average accuracy gap of 0.23. Another surprising phenomenon is that sometimes the student model even shows slightly stronger performance than the teacher model. We conjecture the cause of it may be that our knowledge distillation method cooperates well with purely supervised learning.

Table 1: **Performance gap between teacher and student model.** Experiments are implemented on CIFAR-100, with teacher and student in homogenous architecture. Top-1 accuracy as the evaluation metric. Note that when student model outperforms the teacher model, the gap is negative. The model architecture follows Table 3 in the main paper.

Teacher	72.34	74.31	79.42	75.61	75.61	74.64	75.32 (Avg)
Student (Ours)	72.19	74.11	77.08	76.63	75.35	75.18	75.09 (Avg)
Gap	0.15	0.20	2.34	-1.02	0.26	-0.54	0.23 (Avg)

Combination with White-box Methods Besides serving as a black-box knowledge distillation method, our method can also be combined with existing white-box methods when the teacher model is accessible. In Table 2, we integrate our method with existing white-box methods and show the corresponding results. Our method brings about obvious improvements to RKD Park et al. (2019) (71.73 to 75.32) and steadily pushes ReviewKD Chen et al. (2021), a readily strong white-box method, to a higher level.

Table 2: **Combination with White-box Methods.** Experiments are implemented on CIFAR-100, with teacher and student in homogenous architecture. Top-1 accuracy as the evaluation metric. The model architecture follows Table 3 in the main paper.

RKD Park et al. (2019)	69.61	71.82	71.90	73.35	72.22	71.48	71.73 (Avg)
+ Ours	72.34	74.01	77.38	76.89	75.30	75.90	75.32 (Avg)
ReviewKD Chen et al. (2021)	71.89	73.89	75.63	76.12	75.09	74.84	74.58 (Avg)
+ Ours	72.83	74.52	78.01	77.54	76.21	75.69	75.80 (Avg)

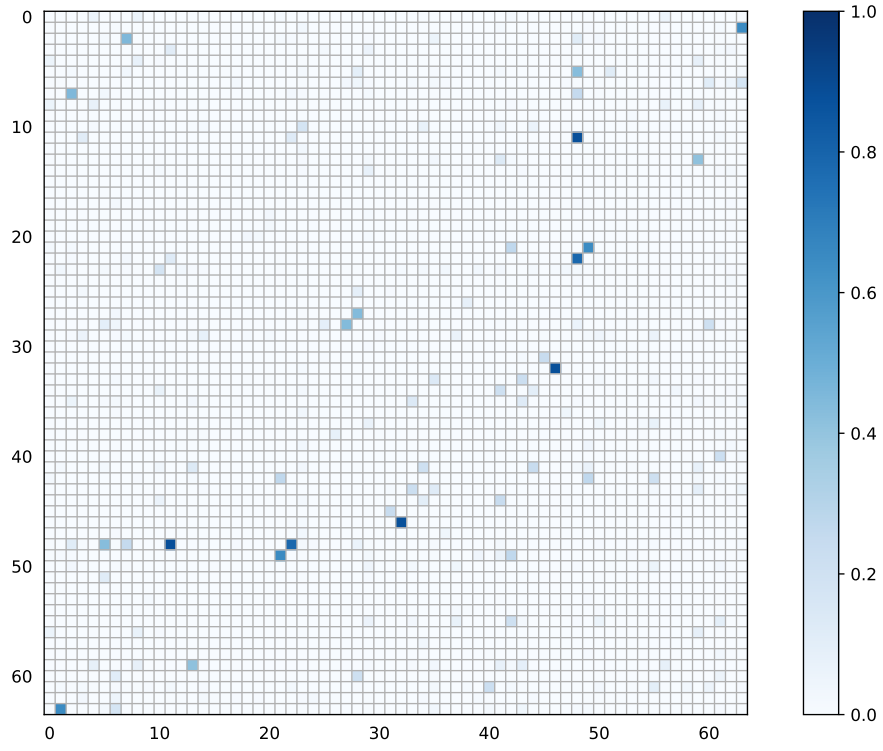
4 MORE VISUALIZATION RESULTS

Input correlation In our method, we enforce the student model to absorb the input correlation knowledge from the teacher model. Here, we visualize the distance between the input correlation matrices of the teacher and student model. The diagonal values are removed for a clearer demonstration. We take ResNet32x4 as the teacher model and ResNet8x4 as the student model and train them on CIFAR-100 Krizhevsky et al. (2009). We calculate the distance on a batch of data with a batch size of 64. As shown in Figure 4, in our method, the student model learns input correlation knowledge from the teacher model better.

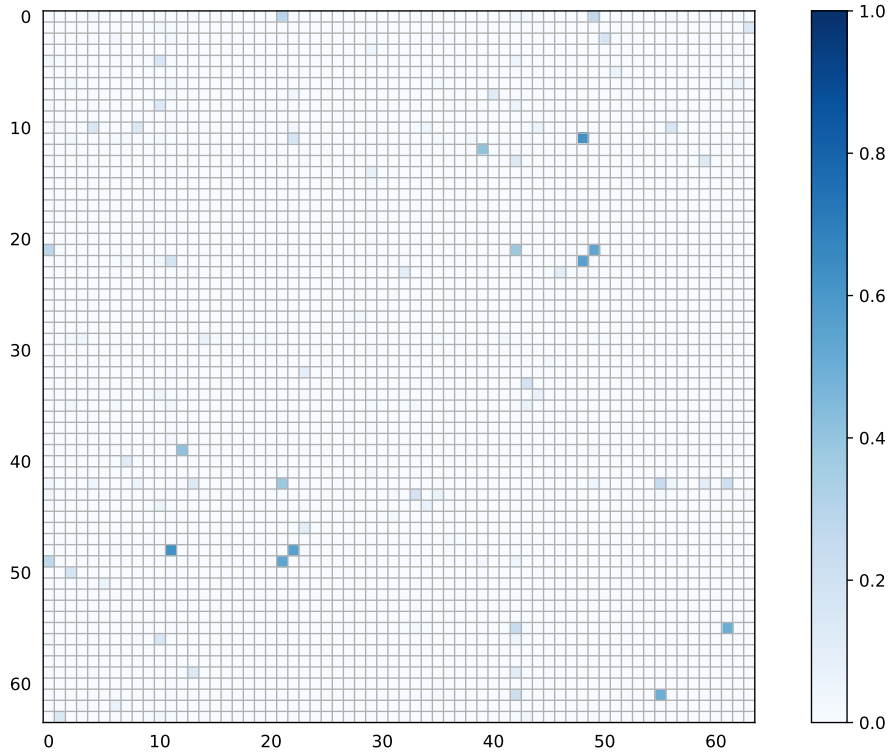
Category Correlation Similarly, the student model mimics the category correlation from the teacher model. We visualize the distance of the category correlation matrix between the teacher and student model. The diagonal values are removed for a clearer demonstration. We take ResNet32x4 as the teacher model and ResNet8x4 as the student model and train them on CIFAR-100 Krizhevsky et al. (2009). We can see from Figure 2(b) that when compared with the vanilla KD Hinton et al. (2015) method (Figure 2(a)), our method reduces the distance of category correlation matrices effectively.

REFERENCES

- Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *arXiv:1503.02531*, 2015.
- Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.
- Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015.

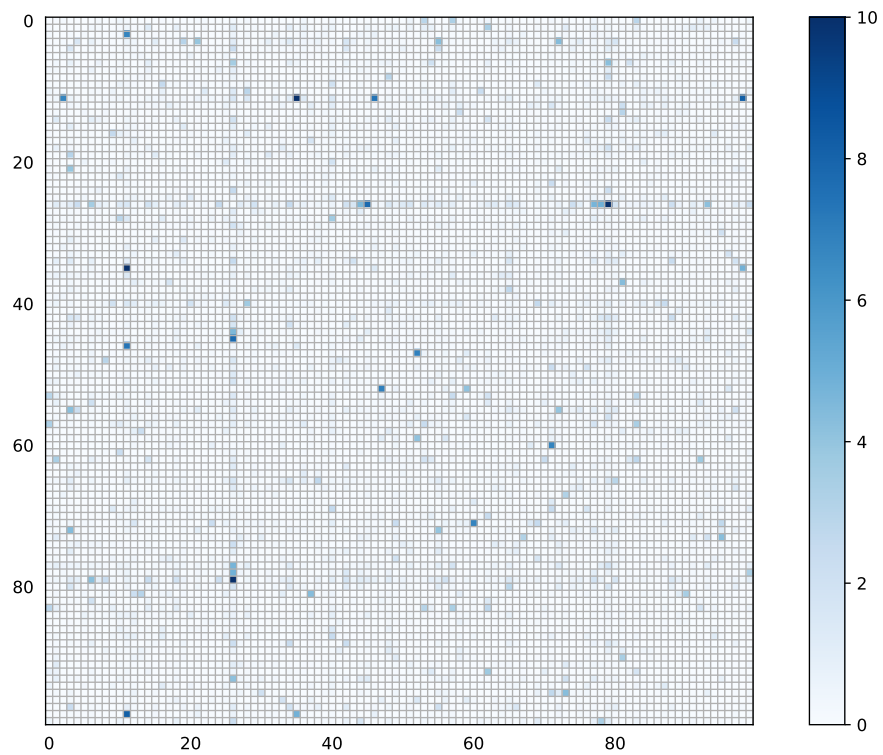


(a) KD Hinton et al. (2015)

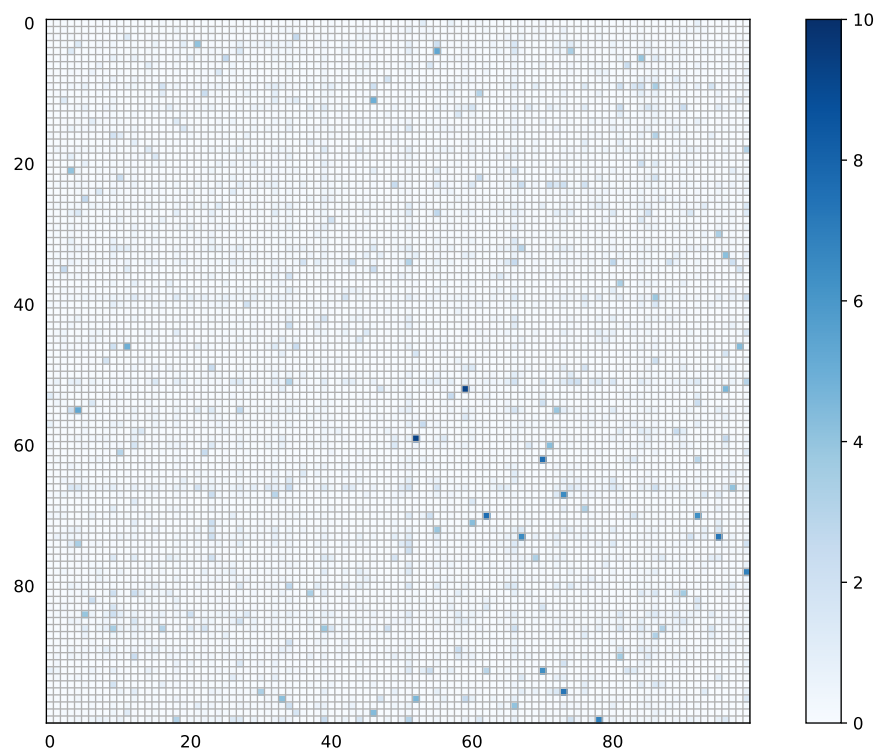


(b) Ours

Figure 1: **Distance of the input correlation matrix** between the teacher and student model. We take ResNet32x4 as the teacher model and ResNet8x4 as the student model and train them on CIFAR-100. We calculate the input correlation matrix on a batch of data with a batch size of 64.



(a) KD Hinton et al. (2015)



(b) Ours

Figure 2: **Distance of the category correlation matrix** between the teacher and student model. We take ResNet32x4 as the teacher model and ResNet8x4 as the student model and train them on CIFAR-100.