Supplementary Materials for: Multi-level Logit Distillation

Ying Jin[1] Jiaqi Wang[2✉*] Dahua Lin[1,2]

[1]CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong
[2]Shanghai AI Laboratory

{jy021,dhlin}@ie.cuhk.edu.hk, wjqdev@gmail.com

1. More Visualization Results

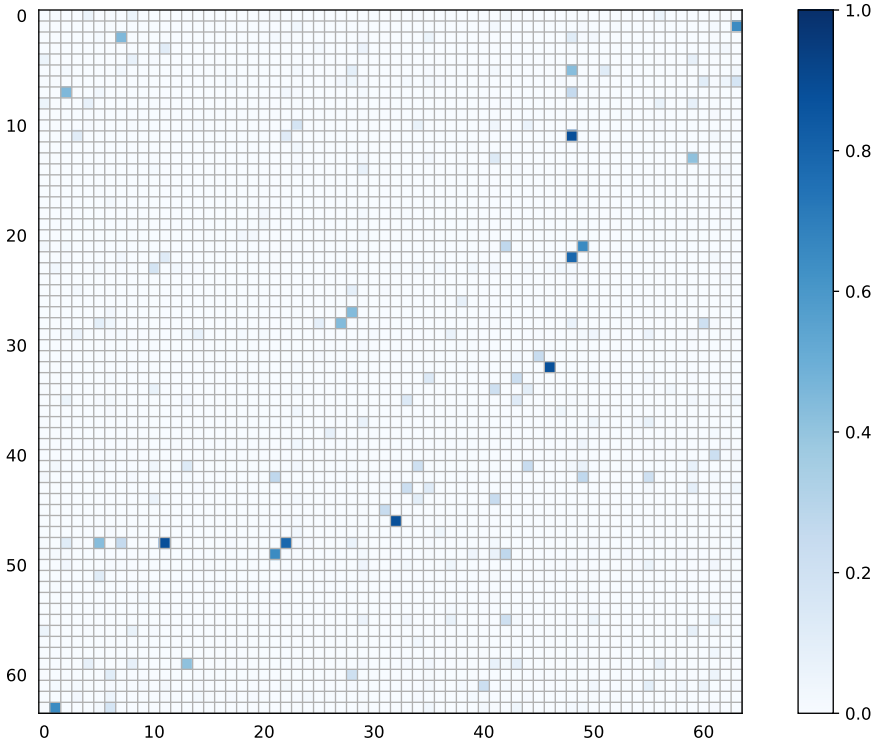
Here, we show clearer figures to visualize the input correlation matrices and category correlation matrices. We take ResNet32x4 as the teacher model and ResNet8x4 as the student model and train them on CIFAR-100 [2]. We calculate the input correlation distance on a batch of data with a batch size of 64. We can see from figures that when compared with the vanilla KD [1] method , our method reduces the distance of both input correlation and category correlation matrices effectively.

References

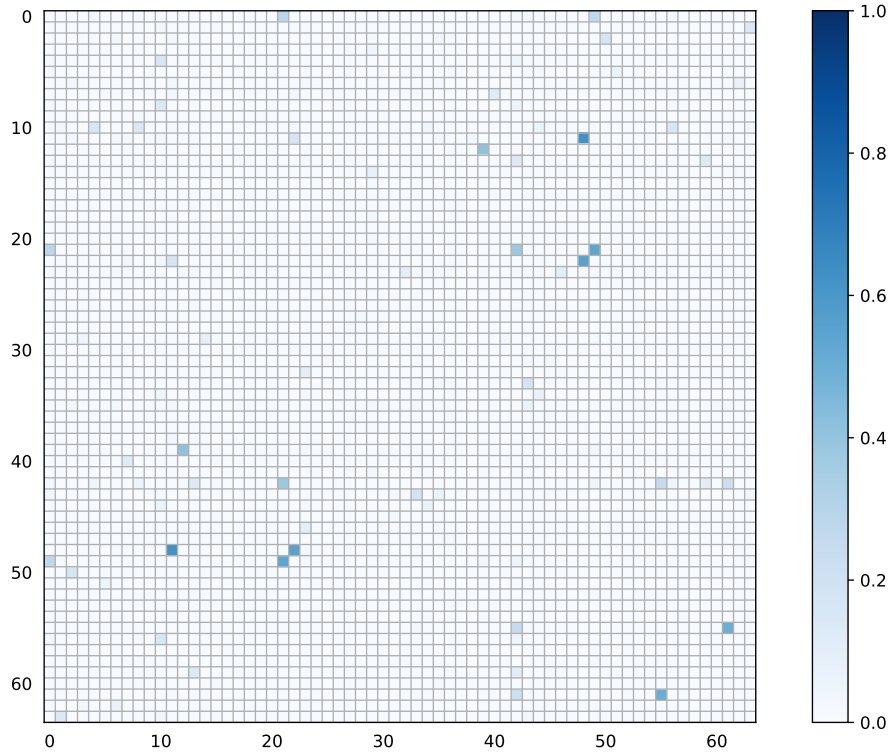
[1] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *arXiv:1503.02531*, 2015. 1, 2, 3

[2] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

*✉Corresponding author.

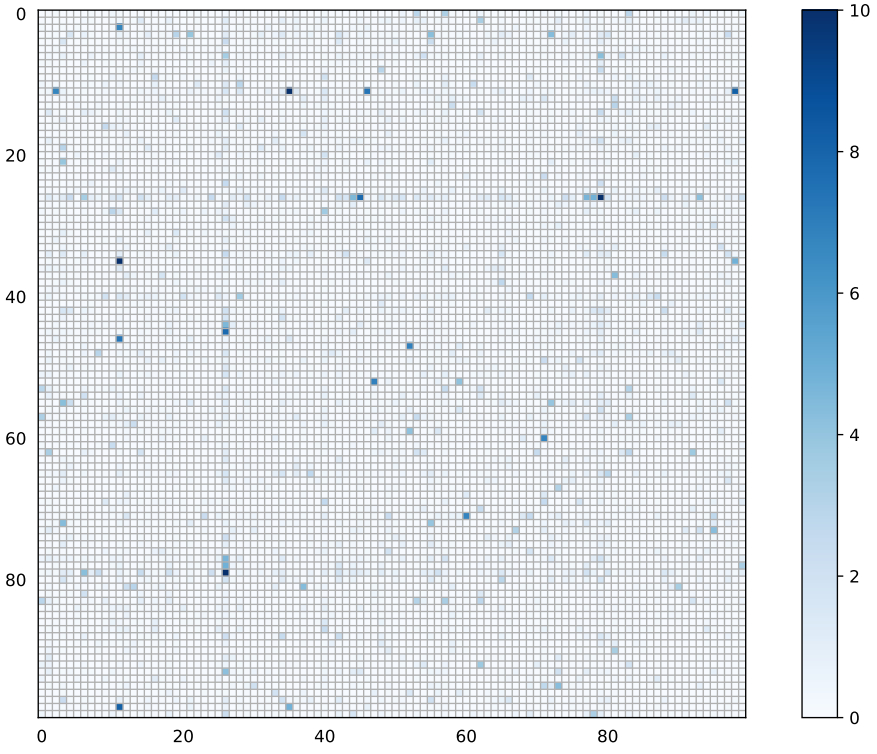


((a)) KD [1]

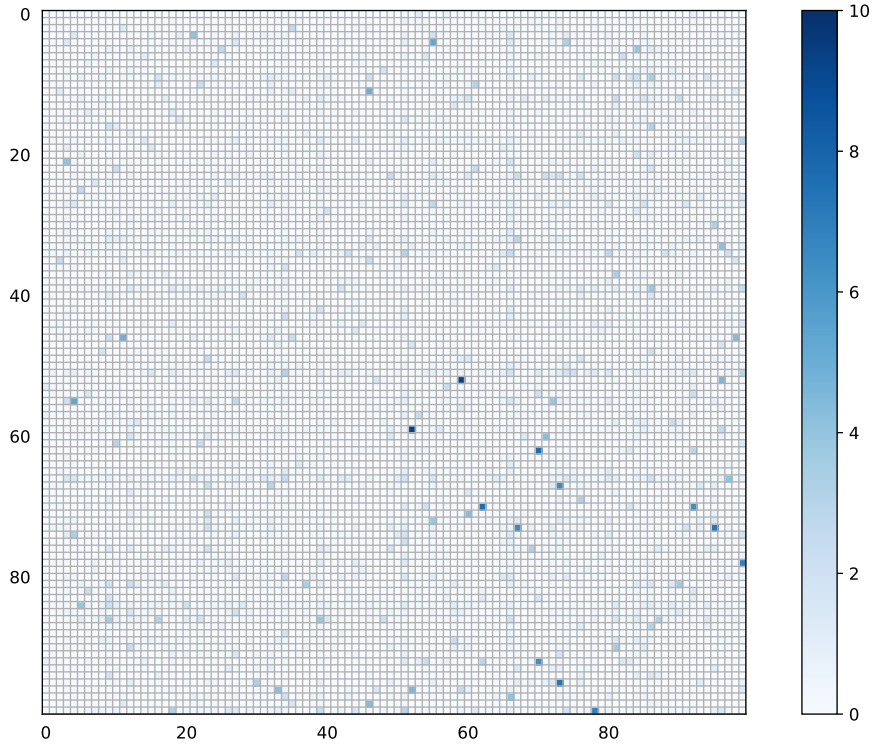


((b)) Ours

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 [1]



((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.