Supplementary Materials for *Self-Regulated Feature Learning via Teacher-free Feature Distillation*

Lujun Li[0000−0002−4329−2707]

Chinese Academy of Sciences, Beijing, China
lilujunai@gmail.com

1 More Discussions

More details about the teacher model exploration experiment. Table 1 presents more detailed results of the teacher model explorations. We use ResNet-110 as a teacher and ResNet-20 as a student. Moreover, student models are trained under the online and offline training paradigm using the same training settings as the main experiment (see detailed experimental settings on CIFAR-100). The results demonstrate that all teacher models achieve performance gains. High-capability teacher models and online training are beneficial for enhancing the performance of student models.

Table 1: Different teacher models for feature distillation of ResNet-20 on CIFAR-100. Auxiliary branches denote the Conv3_x, Conv4_x and FC layers of ResNet-20. We report top-1 mean accuracies (%) over 3 runs.

Teacher-model	Paradigm	Top-1	Gain
ResNet-110	Offline	70.72	1.66
ResNet-20	Offline	70.42	1.36
ResNet-110	Online	70.94	1.88
ResNet-20	Online	70.48	1.42
Auxiliary branches	Online	70.04	0.98

More discussion about feature regularization. The Tf-FD minimizes distillation losses to convey privileged knowledge channel- and layer-wise. There are still significant differences between Tf-FD, and the conventional feature regularization approaches in terms of implementation and feature noise, although we explain why Tf-FD works from regularization. Compared to random noise, feature noise in Tf-FD has richer visual information and fewer semantic gaps. Additionally, there are various distillation loss options and methods for balancing loss weights for simple adjustment of the regularization effect. For some models on CIFAR-100 and ImageNet, Tf-FD can perform better than conventional regularization techniques.

More analysis of feature salience metrics. Inspired by the model filter pruning, we also evaluate different salience metrics in pruning methods for intra Tf-FD, including extraction of salient features based on l_p norm ($p = 2$), entropy [1], BN scaling factor [7] and HRank [4]. Our experiment results illustrate that entropy and HRank obtain slightly more gains than simple l_p-norm, and BN scaling factor achieves similar performance. Contrary to pruning methods that remove filters/weights of redundant features after feature ranking, intra-layer Tf-FD uses salient features to distill redundant ones without altering the network structure.

1.1 More Experimental Settings

Detailed experimental settings on CIFAR-100. The CIFAR-100 dataset is used for the trials without any further data augmentation. The weight decay is 5×10^{-4}, and the optimizer is SGD. We employ a warm-up for weight reduction in the first 20 epochs following ReviewKD [5]. We refer to the unified teacher setting of CRD for the execution of additional distillation and Tf-FD†. Dropblock [2] is applied to the output of the first two groups for various regulation methods. We are applying Dropout [3] to the output of the previous group.
Detailed experimental settings on ImageNet. ResNet18 is trained using 100 training epochs in ImageNet experiments, which is the standard setting for distillation. Following Dropblock [2], label smoothing [6] is applied to the outputted logits. We perform inter Tf-FD for the neighboring layers to simplify the computations. Warm-up and early-decay schedules are conducted for the weight of losses.

References

1. Fanxu Meng, Hao Cheng, K.L.Z.X.R.J.X.S., Lu, G.: Filter grafting for deep neural networks. In: CVPR (2020) 2
2. Ghiasi, G., Lin, T.Y., Le, Q.V.: Dropblock: A regularization method for convolutional networks. In: NeurIPS (2018) 2
3. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NeurIPS (2012) 2
4. Lin, M., Ji, R., Wang, Y., Zhang, Y., Zhang, B., Tian, Y., Shao, L.: Hrank: Filter pruning using high-rank feature map. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1529–1538 (2020) 2
5. Pengguang Chen, Shu Liu, H.Z., Jia, J.: Distilling knowledge via knowledge review. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 2
6. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: ICCV (2016) 2
7. Zhuang Liu, Jianguo Li, Z.S.G.H.S.Y., Zhang, C.: Learning efficient convolutional networks through network slimming. In: ICCV (2017) 2