DearKD: Data-Efficient Early Knowledge Distillation for Vision Transformers Supplement Materials

Xianing Chen[1]*, Qiong Cao[2]†, Yujie Zhong[3], Jing Zhang[4], Shenghua Gao[156]†, Dacheng Tao[24]
[1]ShanghaiTech University, [2]JD Explore Academy, [3]Meituan Inc.,
[4]The University of Sydney, [5]Shanghai Engineering Research Center of Intelligent Vision and Imaging,
[6]Shanghai Engineering Research Center of Energy Efficient and Custom AI IC
{chenxn1,gaoshh}@shanghaitech.edu.cn {mathqiong2012,dacheng.tao}@gmail.com
jaszhong@hotmail.com jing.zhang1@sydney.edu.au

1. The image regularization term of DF-DearKD

The image regularization term $R(\cdot)$ consists of two terms: the prior term R_{prior} [2] that acts on image priors and the BN regularization term R_{BN} that regularizes feature map distributions:

$$R(x) = R_{\text{prior}}(x) + R_{\text{BN}}(x) \tag{1}$$

Specifically, R_{prior} penalizes the total variance and l2 norm of x, respectively.

$$\mathcal{R}_{\text{prior}}(x) = \alpha_{TV}\mathcal{R}_{TV}(x) + \alpha_{l_2}\mathcal{R}_{l_2}(x) \tag{2}$$

$\mathcal{R}_{\text{BN}}$ matches the feature statistics, i.e., channel-wise mean $\mu(x)$ and variance $\sigma^2(x)$ of the current batch to those cached in the BN [1] layers at all levels:

$$\mathcal{R}_{BN}(x) = \alpha_{BN}\sum_{l=1}^{L}\left\|\mu_l(x) - \mu_l^{BN}\right\|_2 + \left\|\sigma_l^2(x) - \sigma_l^{2BN}\right\|_2 \tag{3}$$

where L is the total number of BN layers.

2. Generated samples from DF-DearKD

Figure 1 shows samples generated by our method from an ImageNet-pretrained RegNetY-16GF model. Remarkably, given just the pre-trained teacher model, we observe that our method is able to generate images with high fidelity and resolution.

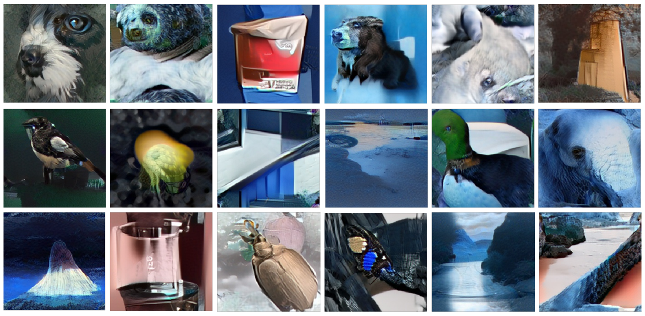

Figure 1. **Images generated by our method on RegNetY-16GF model pre-trained with ImageNet.**

Epochs number	200	225	250	275	300
Accuracy	74.3	74.6	74.8	74.7	74.6

Table 1. **Ablation of different epochs number of the first stage of DearKD evaluated on ImageNet classification.** DearKD-Ti is used.

3. Analysis of the number of epochs for the first stage of DearKD

In this section, we ablate the number of epochs for the first stage of our DearKD. As can be seen in Table 1, training the model in the first stage with 250 epochs achieves the best 74.8% Top-1 accuracy among other settings. It is not surprising that training the model in the first stage with less epochs will lead to worse performance. But, for models trained with 300 epochs, the inductive biases knowledge from CNNs are not saturated. So, we use Equation (6) in the second stage except that we set β to 0 and let α linearly increase to 1. Besides, for models trained with 1000 epochs, we empirically select 800 as the number of epochs for the first stage.

*This work was done when Xianing Chen was intern at JD Explore Academy.

†Corresponding authors.

4. More implement details of DF-DearKD

We filter out ambiguous images whose output logits from a pre-trained ResNet-101 are less than 0.1 and finally synthesize 600k images to train our transformer student network from scratch. Then, we use the target label for inversing the RegNetY-16GF as our ground truth. The RegNetY-16GF can achieve 100% accuracy on the generated samples. This phenomenon is the same as that in [4]. So, we use a pre-trained ResNet-101 from pytorch [3] that achieves 77.37% top-1 accuracy on ImageNet as our teacher model, which can provide good results as well as inductive biases clues. We use AdamW optimizer with learning rate 0.0005 and cosine learning scheduler. The model is trained from scratch for 1000 epochs. A batch size of 1024 is used. We train the model in the first stage with 800 epochs. We use Mixup [6], Cutmix [5], Random Erasing [7] and Random Augmentation [7] for data augmentation. Experiments are conducted on 4 NVIDIA TESLA V100 GPUs.

5. Limitation and Future works

Although DF-DearKD can generate high quality images, it still has difficulty in handling human-related classes due to the limited information stored in the feature statistics. Moreover, we generate lots of samples which takes a lot of time and computation costs even we do not use any real images. There is still a gap between training with generated samples and real images. In the future, we plan to investigate more in model inversion or image generation to further improve training data quality and diversity.

Besides, to further explore the data efficiency of training vision transformers under different settings (i.e. full ImageNet, partial ImageNet and data-free case), we plan to distill other kinds of IBs for transformers and investigate how to introduce transformers' intrinsic IBs in the future study. The data-free setting would be a particularly interesting case to cope with the emerging concern of data privacy in practice.

References

[1] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 1

[2] A. Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. 2015. 1

[3] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 2

[4] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8715–8724, 2020. 2

[5] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision*. 2

[6] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. 2017. 2

[7] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7), 2017. 2