Appendix

Xueqing Deng*[1,3], Dawei Sun*[2], Shawn Newsam[3], and Peng Wang[1]

[1] ByteDance Inc. {xueqingdeng,peng.wang}@bytedance.com
[2] ECE, UIUC, {daweis2}@illinois.edu
[3] EECS, UC Merced, {snewsam}@ucmerced.edu

A Implementation Details on Transformer

In this section, we present the implementation details on the experiments performed on transformer. We select ViT-B [2] with patch size of 16 as our teacher model and DeiT-Tiny [4] as our student model. We reproduce the baseline result with 4 GPUs and the total batch size is 1024. However, for searching the distillation process, we have to reduce the batch size to 256 due to limited GPU memory as we have pathways between the feature maps from teacher and student. Meanwhile, we keep the same batch size for retraining after searching. The most significant difference between the implementations of convolutional neural networks (CNNs) and transformers is the transform block. Our experimental results show that the proposed transform block on CNNs is not applicable to transformer yielding much worse performance on distillation compared to non-distillation. Therefore, we employ a transformer-style block to serve as a transform block for feature transfer between the teacher and student whose architectures are transformers as shown in Fig. 1. We follow similar search pipeline with a search learning rate of 1e-3. Once the distillation process is obtained, we train the models with 150 epochs for both ReviewKD [1] and our proposed `DistPro` following the same configurations in DeiT [4].

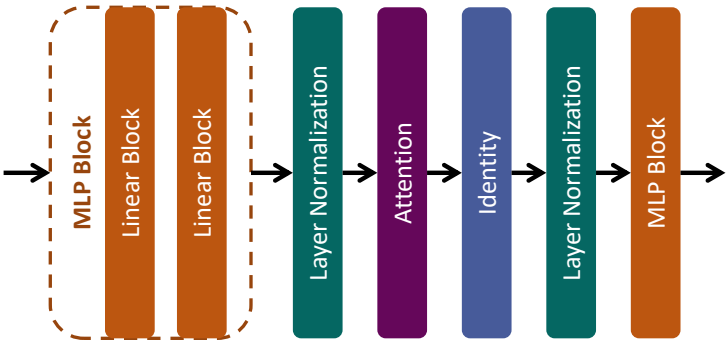

Fig. 1. Transform block architecture for transformer. The linear block is a linear layer followed with GELU activation and dropout.

Setting	Search teacher	Search student	Retrain teacher	Retran student	Top-1 (%)
(b)	ResNet34	ResNet18	ResNet34	ResNet18	71.89
	ResNet50	MobileNet	ResNet34	ResNet18	71.87
	ResNet50	MobileNet	ResNet50	MobileNet	73.26
	ResNet34	ResNet18	ResNet50	MobileNet	73.22

Table 1. More results of transferring searched $\mathcal{A}$ cross search-retrain networks performed on ImageNet1K with 100 epochs. Top-1 accuracy on validation set is reported.

B More Experimental Results on Transferable Process

In this section, we provide more experimental results to analyze how the learned process $\mathcal{A}$ affects transferable distillation. As shown in the paper, for setting (b), we adopt the searched process with student/teacher networks of MobileNet/ ResNet50 to networks of ResNet18/ResNet34 since they have same feature pathways at similar corresponding layers. We conduct the reversed testing as well. As shown, all the results are closed, demonstrating the process could be transferred when similar pathways exist.

C More Dense Prediction Task: Object Detection

In this section, we present the results on another dense prediction task: detection on COCO dataset [3]. We follow the same distillation configuration as in [1]. Similarly, we distill the backbone output features from the teacher and student. Different from the other dense prediction tasks in the paper, searching performed on detection is complicated due to the complex detection head and neck. We performed our experiments of object detection using the searched process from ImageNet1K with the teacher network ResNet50 and student network MobileNetV2. We follow all the other training configurations from ReviewKD [1]. As shown in Tab. 2, our proposed `DistPro` outperforms the baseline method ReviewKD.

	Method	mAP	AP50	AP70	APl	APm	APs
Teacher	Mask R-CNN w/ ResNet50-FPN	37.17	58.60	39.88	53.30	39.49	18.63
Student	Mask R-CNN w/ MobileNetV2-FPN	28.37	47.19	29.95	41.70	29.01	12.09
	+ReviewKD	31.56	50.70	33.44	47.39	32.44	12.76
	+DistPro (ours)	**32.37**	**51.22**	**34.02**	**48.16**	**33.29**	**13.17**

Table 2. Quantitative results on object detection performed on COCO.

D Trade-off between distillation efficiency and accuracy

As we mentioned in the paper, our proposed method can contribute to fast distillation. In this section, we discuss the trade-off between the training speed and accuracy. We conduct the experiments by setting different thresholds to drop

the alphas during distillation. Tab. 3 shows the results. As shown in the table, when we increase the drop rate from 0.3 to 0.7, more alphas will be discarded. This can lead to faster distillation speed, since less distillation loss need to be computed but result in lower network accuracy. Keeping around 50% α remains the best performance regarding speed and accuracy.

Threshold	Remain alphas (%)	Computation Cost (GPU hours) ↓	Top-1 accuracy (%) ↑
0.3	80	102	73.25
0.5	50	81	73.26
0.6	40	73	73.22
0.7	26	60	72.89

Table 3. Trade-off between the distillation cost and accuracy. Experiments are performed on ImageNet1K with teacher network of ResNet50 and student network of MobileNetV2.

E Analysis on Searched Process

In this section, we provide more analysis on how the searched process affects the distillation procedure. Fig. 2 and 3 show the visualization of the searched process $\mathcal{A}$. The values of α in the figure are normalized. We show three groups of $\mathcal{A}$ learned with combinations of different networks and lr schedulers. The first comparison is made between different lr schedulers during searching (left and middle columns). The left column shows the $\mathcal{A}$ learned with a step lr scheduler and the middle one shows that with cosine scheduler. It is noted that the lr scheduler is only used for optimizing the network weights w instead of α in equation (2). We can see even though the lr scheduler is different which affects the training loss, the process of α remains similar (the curves show closed shapes). However, when we change the networks while keeping lr scheduler the same (middle and right columns), this will result in significant change in the values of α. Even though the curve shapes are different, the behavior are similar. For example, the low-level feature maps from teacher to student at the beginning remain high importance and later the importance decreases. Later, the high-level feature maps will raise the importance during distillation. In details, the importance of connect 5-3, 5-4 and 5-5 in total has large percentage for both networks, indicating the high-level information from teacher network becomes more and more significant as the training step increases.

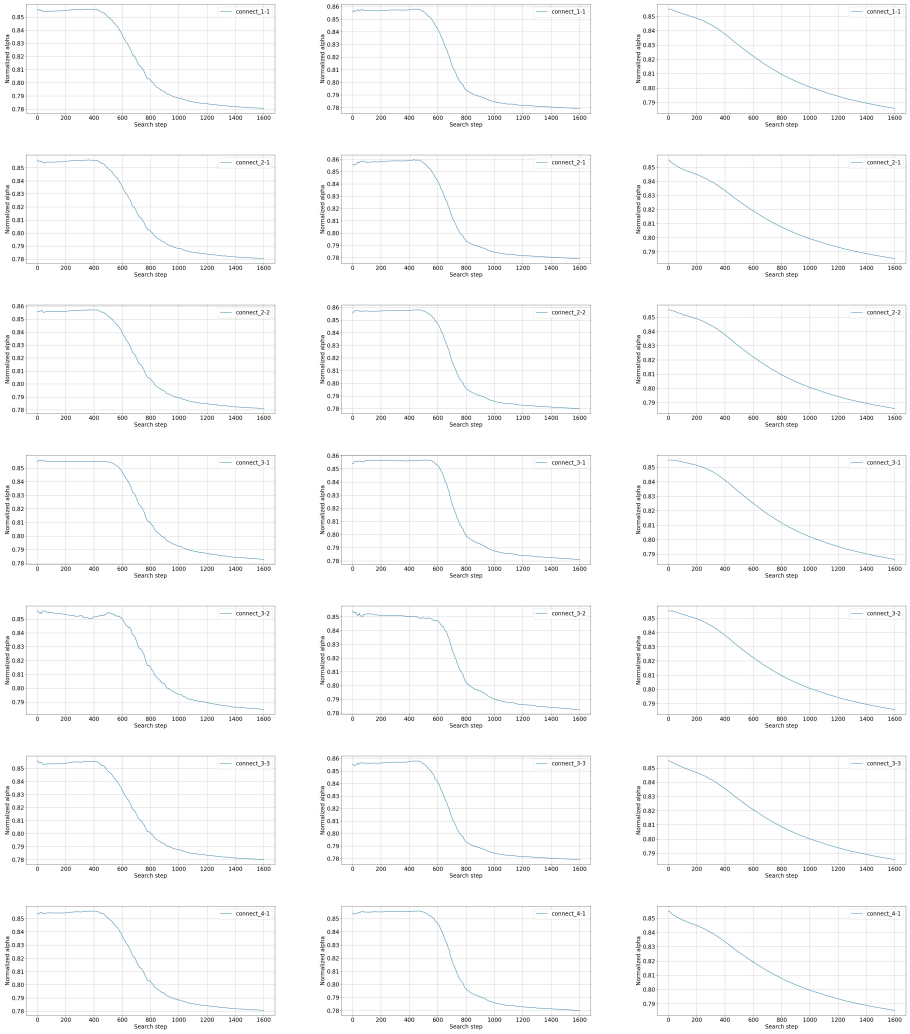

Fig. 2. Visualization of searched distillation process performed on ImageNet1K $\mathcal{A} = \{\alpha_t^{i,j}\}, 0 \leq t \leq T_{search}, 1 \leq i \leq \mathcal{C}_T, 1 \leq j \leq i$, where i denotes the feature map from teacher network, while j denotes that from student network. The searching configuration from left to right is: Resnet50 to MobilenetV2 with **step** lr scheduler, Resnet50 to MobilenetV2 with **cosine** lr scheduler, Resnet34 to ResNet18 with **cosine** lr scheduler. The remaining connections are shown in the next figure.

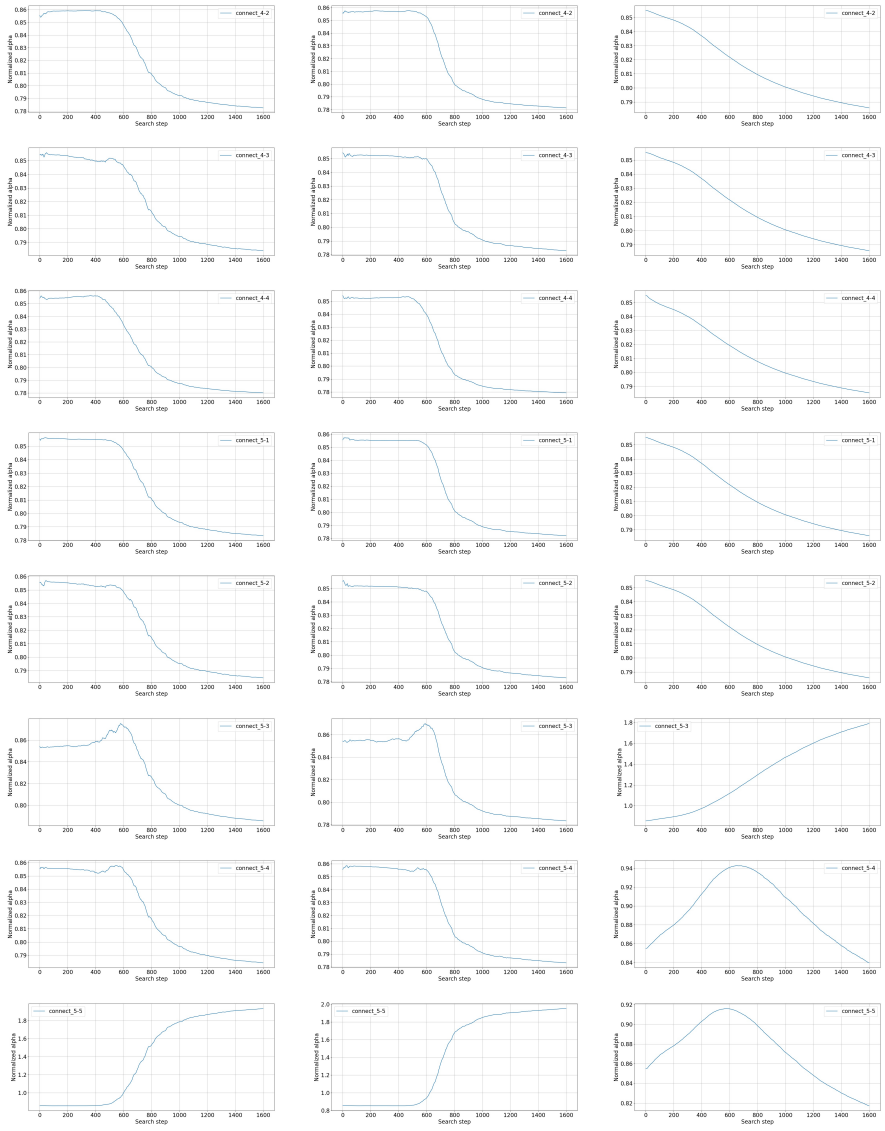

Fig. 3. The remaining visualization of searched distillation process performed on ImageNet1K followed by previous figure. The searching configuration from left to right is: Resnet50 to MobilenetV2 with **step** lr scheduler, Resnet50 to MobilenetV2 with **cosine** lr scheduler, Resnet34 to ResNet18 with **cosine** lr scheduler.

References

1. Chen, P., Liu, S., Zhao, H., Jia, J.: Distilling knowledge via knowledge review. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5008–5017 (2021)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
4. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021)