Supplementary Materials: A Fast Knowledge Distillation Framework for Visual Recognition

Zhiqiang Shen[1,2,3] and Eric Xing[1,3]

[1] Carnegie Mellon University, Pittsburgh, USA
[2] Hong Kong University of Science and Technology, Hong Kong, China
[3] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
zhiqiangshen@cse.ust.hk,epxing@cs.cmu.edu
Project Page: http://zhiqiangshen.com/projects/FKD/index.html

Appendix

A Visualization, Analysis and Discussion

To investigate the learned differences of information between ReLabel and FKD, we depict the intermediate attention maps using gradient-based localization [3]. There are three important observations that align our aforementioned analyses in Fig. 1 and 2.

(i) FKD's predictions are less confident than ReLabel with more surrounding context; This is reasonable since in random-crop training, many crops are basically backgrounds (context), the soft predicted label from the teacher model might be completely different from the ground-truth one-hot label and the training mechanism of FKD can leverage the additional information from context.

(ii) FKD's attention maps have a larger active area on the object regions, which indicates that FKD trained model utilizes more cues for prediction and also captures more subtle and fine-grained information. However, it is interesting to see that the *guided backprop* is more focused than ReLabel.

(iii) ReLabel's attention is more aligned with PyTorch pre-trained model, while FKD's results are substantially unique to them. It implies that FKD's learned attention differs significantly from one-hot and global label map learned models.

B Training Details and Experimental Settings

Training details for Table 3 of the main text. We employ the training settings and hyper-parameters following Table 1, which are the same as ReLabel. We use 4 as the number of crops in each image during training.
Training details for Table 5 of the main text. When comparing our FKD with ViT [1]/DeiT [6]/SReT [4] (Table 5 of the main text), we employ the training settings and hyper-parameters following Table 2.

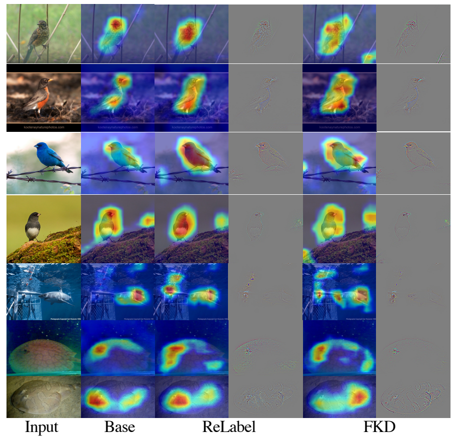

Fig. 1. Visualization of learned attention map using GradCAM [3,2]. "Base" indicates the pre-trained PyTorch model. In each group of ReLabel and FKD, left is *Grad-CAM* and right is *Guided Backprop*.

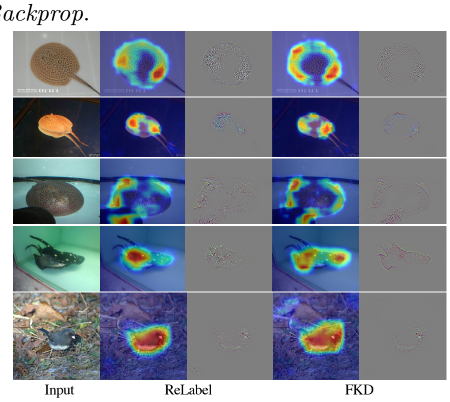

Fig. 2. More visualization of response/attention maps.

Training details for Table 8 of the main text. The training settings and hyper-parameters of FKD with FBNet-C100 [7] and EfficientNetv2-B0 [5] backbones (Table 8 of the main text) are provided in Table 2 which are the same as the training protocol on ViT, DeiT and SReT. We use 4 as the number of crops in each image during training.

Table 1. Training hyper-parameters and details for ReLabel [8] and FKD used in Table 3 of the main text.

Method	ReLabel [8] or FKD
Teacher	EfficientNet-L2-ns-475
Epoch	300
Batch size	1,024
Optimizer	SGD
Init. *lr*	0.1
lr scheduler	cosine
Weight decay	1e-4
Random crop	Yes
Flipping	Yes
Warmup epochs	5
Color jittering	Yes

Table 2. Training hyper-parameters and details for the comparison in Table 5 of the main text when employing ViT [1], DeiT [6] and SReT [4] as the backbone networks. Table is adapted from [6].

Method	ViT-B [1]	DeiT [6]/SReT [4]	FKD
Epoch	300	300	300
Batch size	4096	1024	1024
Optimizer	AdamW	AdamW	AdamW
Init. *lr*	0.003	0.001	0.002
lr scheduler	cosine	cosine	cosine
Weight decay	0.3	0.05	0.05
Warmup epochs	3.4	5	5
Label smoothing	None	0.1	None
Dropout	0.1	None	None
Stoch. Depth	None	0.1	0.1
Repeated Aug	None	Yes	None
Gradient Clip.	Yes	None	None
Rand Augment	None	9/0.5	None
Mixup prob.	None	0.8	None
Cutmix prob.	None	1.0	None
Erasing prob.	None	0.25	None

References

1. 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. In: International Conference on Learning Representations (2020) 1, 3
2. Gildenblat, J., contributors: Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam (2021) 2
3. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017) 1, 2
4. Shen, Z., Liu, Z., Xing, E.: Sliced recursive transformer. arXiv preprint arXiv: 2111.05297 (2021) 1, 3
5. Tan, M., Le, Q.: Efficientnetv2: Smaller models and faster training. In: International Conference on Machine Learning. pp. 10096–10106. PMLR (2021) 1
6. 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) 1, 3
7. Wu, B., Dai, X., Zhang, P., Wang, Y., Sun, F., Wu, Y., Tian, Y., Vajda, P., Jia, Y., Keutzer, K.: Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10734–10742 (2019) 1
8. Yun, S., Oh, S.J., Heo, B., Han, D., Choe, J., Chun, S.: Re-labeling imagenet: from single to multi-labels, from global to localized labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2340–2350 (2021) 3