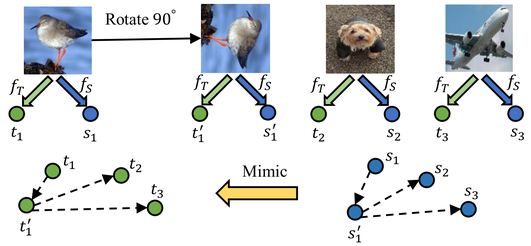
Hierarchical Self-supervised Augmented Knowledge Distillation

Chuanguang Yang[1,2], Zhulin An[1], Linhang Cai[1,2] and Yongjun Xu[1]
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences

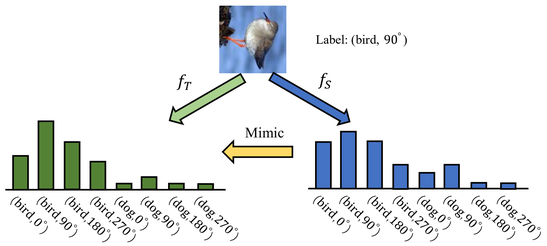
IJCAI 2021 MONTREAL

Introduction

- The current pattern of KD can be summarized as two critical aspects: (1) what kind of knowledge encapsulated in teacher network can be explored for KD; (2) How to effectively transfer knowledge from teacher to student.
- We introduce a self-supervised augmented distribution that encapsulates the unified knowledge of the original classification task and auxiliary self-supervised task [1] as the richer dark knowledge for the field of KD.



a) Self-supervised contrastive relationship [2].



b) Our proposed self-supervised augmented distribution.

Compared with the previous SOTA SSKD [2], our method can effectively learn knowledge from self-supervised representation learning without interfering with the original fully-supervised classification task.

- We propose a one-to-one probabilistic knowledge distillation framework by leveraging the architectural auxiliary classifiers, facilitating comprehensive knowledge transfer and alleviating the mismatch problem of abstraction levels when existing a large architecture gap.
- HSAKD significantly refreshes the results achieved by previous SOTA SSKD on standard image classification benchmarks. It can also learn well-general feature representations for downstream semantic recognition tasks.

Reference

[1] Gidaris, Spyros and Singh, Praveer and Komodakis, Nikos. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018
[2] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. In *ECCV*, pages 588–604, 2020.

Methodology

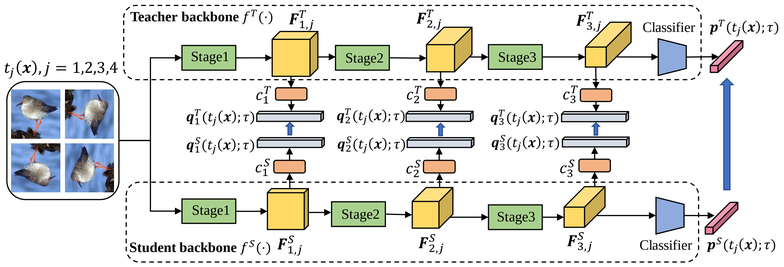


Figure 1. Overview of our proposed HSAKD.

Pre-training a teacher network

Given the self-supervised transformed images , we train the backbone f^T using the conventional labels y and L auxiliary classifiers $\{c_l^T(\cdot)\}_{l=1}^L$ by our proposed self-supervised augmented labels k_j. The overall loss is formulated as $\mathcal{L}_T$:

$$\mathcal{L}_T = \mathbb{E}_{\boldsymbol{x}\in\mathcal{X}}[\mathcal{L}_{ce}(\boldsymbol{p}^T(\boldsymbol{x};\tau), y) + \frac{1}{M}\sum_{j=1}^{M}\sum_{l=1}^{L}\mathcal{L}_{ce}(\boldsymbol{q}_l^T(t_j(\boldsymbol{x});\tau), k_j)] \tag{1}$$

Training a student network supervised by a teacher network

Task loss. Given the input image $\boldsymbol{x}$, we train the backbone f^S using the conventional labels y with the cross-entropy loss as $\mathcal{L}_{task}$:

$$\mathcal{L}_{task} = \mathcal{L}_{ce}(\boldsymbol{p}^S(\boldsymbol{x};\tau), y) \tag{2}$$

Mimicry loss from class probability distributions. Given the self-supervised transformed images $\{t_j(\boldsymbol{x})\}_{j=1}^M$, we guide the the class probability distributions $\boldsymbol{p}^S(t_j(\boldsymbol{x});\tau)$ generated from the backbone f^S to mimic the class probability distributions $\boldsymbol{p}^T(t_j(\boldsymbol{x});\tau)$ generated from the teacher backbone f^T by KL-divergence D_{KL}:

$$\mathcal{L}_{kl\_p} = \frac{1}{M}\sum_{j=1}^{M}\tau^2 D_{\mathrm{KL}}(\boldsymbol{p}^T(t_j(\boldsymbol{x});\tau) \parallel \boldsymbol{p}^S(t_j(\boldsymbol{x});\tau)) \tag{3}$$

Mimicry loss from self-supervised augmented distributions. Given the self-supervised transformed images $\{t_j(\boldsymbol{x})\}_{j=1}^M$, we guide the self-supervised augmented distributions $\{\boldsymbol{q}_l^S(t_j(\boldsymbol{x});\tau)\}_{l=1}^L$ generated from L auxiliary classifiers $\{c_l^S\}_{l=1}^L$ of student network f^S to mimic the corresponding self-supervised augmented distributions $\{\boldsymbol{q}_l^T(t_j(\boldsymbol{x});\tau)\}_{l=1}^L$ generated from L auxiliary classifiers $\{c_l^T\}_{l=1}^L$ of teacher network f^T by KL-divergence D_{KL}:

$$\mathcal{L}_{kl\_q} = \frac{1}{M}\sum_{j=1}^{M}\sum_{l=1}^{L}\tau^2 D_{\mathrm{KL}}(\boldsymbol{q}_l^T(t_j(\boldsymbol{x});\tau) \parallel \boldsymbol{q}_l^S(t_j(\boldsymbol{x});\tau)) \tag{4}$$

Overall loss. We summarize the task loss and mimicry loss as the overall loss $\mathcal{L}_S$:

$$\mathcal{L}_S = \mathbb{E}_{\boldsymbol{x}\in\mathcal{X}}[\mathcal{L}_{task} + \mathcal{L}_{kl\_q} + \mathcal{L}_{kl\_p}] \tag{5}$$

Experiments

Ablation study.

- ***Effect of loss terms (left):*** each loss term is indispensable, and $\mathcal{L}_{kl\_q}$ contributes most by distilling the self-supervised augmented distributions.
- ***Ablation study of auxiliary classifiers (right):*** each auxiliary classifier is indispensable. The auxiliary classifier attached in the deeper layer often achieves more accuracy gains than that in the shallower layer. Using all auxiliary classifiers can maximize accuracy gains.

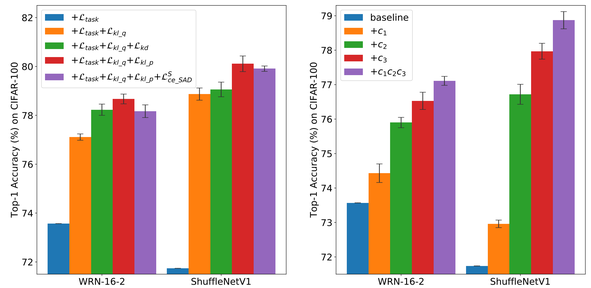


Figure 2. Ablation study of loss terms (***left***) and auxiliary classifiers (***right***) on CIFAR-100.

Results on Image Classification. Our HSAKD significantly outperforms the best-competing method SSKD by an average accuracy gain of 2.56% on CIFAR-100, a top-1 gain of 0.77% on ImageNet, a top-1 gain of 3.63% on STL-10 and a top-1 gain of 3.50% on TinyImageNet.

Table 1: Top-1 accuracy (%) comparison of SOTA distillation methods on CIFAR-100.

Teacher Student	WRN-40-2 WRN-16-2	WRN-40-2 WRN-40-1	ResNet56 ResNet20	ResNet32×4 ResNet8×4	VGG13 MobileNetV2	ResNet50 MobileNetV2	WRN-40-2 ShuffleNetV1	ResNet32×4 ShuffleNetV2
Teacher	76.45	76.45	73.44	79.63	74.64	79.34	76.45	79.63
Teacher*	80.70	80.70	77.20	83.73	78.48	83.85	80.70	83.73
Student	73.57(±0.23)	71.95(±0.59)	69.62(±0.26)	72.95(±0.24)	73.51(±0.26)	73.51(±0.26)	71.74(±0.35)	72.96(±0.33)
KD	75.23(±0.23)	73.90(±0.44)	70.91(±0.10)	73.54(±0.26)	75.21(±0.24)	75.80(±0.46)	75.83(±0.18)	75.43(±0.33)
FitNet	75.30(±0.42)	74.30(±0.42)	71.21(±0.16)	75.37(±0.12)	75.42(±0.34)	75.41(±0.07)	76.27(±0.18)	76.91(±0.06)
AT	75.64(±0.31)	74.32(±0.23)	71.35(±0.09)	75.06(±0.19)	74.08(±0.21)	76.57(±0.20)	76.51(±0.44)	76.32(±0.12)
AB	71.26(±1.32)	74.55(±0.46)	71.56(±0.19)	74.31(±0.09)	74.98(±0.44)	75.87(±0.39)	76.43(±0.09)	76.40(±0.29)
VID	75.31(±0.22)	74.23(±0.28)	71.35(±0.09)	75.07(±0.35)	75.67(±0.13)	75.97(±0.08)	76.24(±0.44)	75.98(±0.41)
RKD	75.33(±0.14)	73.90(±0.26)	71.67(±0.08)	74.17(±0.22)	75.54(±0.36)	76.20(±0.06)	75.74(±0.32)	75.42(±0.25)
SP	74.35(±0.59)	72.91(±0.24)	71.45(±0.38)	75.44(±0.11)	75.68(±0.35)	76.35(±0.14)	76.40(±0.37)	76.43(±0.21)
CC	75.30(±0.03)	74.46(±0.05)	71.44(±0.10)	74.40(±0.24)	75.66(±0.33)	76.05(±0.25)	75.63(±0.30)	75.74(±0.18)
CRD	75.81(±0.33)	74.76(±0.25)	71.83(±0.42)	75.77(±0.24)	76.13(±0.16)	76.89(±0.27)	76.37(±0.23)	76.51(±0.09)
SSKD	76.16(±0.17)	75.84(±0.04)	70.80(±0.02)	75.83(±0.29)	76.21(±0.16)	78.21(±0.16)	76.71(±0.31)	77.64(±0.24)
Ours	77.20(±0.17)	77.00(±0.21)	72.58(±0.33)	77.26(±0.14)	77.45(±0.21)	78.79(±0.11)	78.51(±0.20)	79.93(±0.11)
Ours*	**78.67**(±0.20)	**78.12**(±0.25)	**73.73**(±0.10)	**77.69**(±0.05)	**79.27**(±0.12)	**79.43**(±0.24)	**80.11**(±0.32)	**80.86**(±0.15)

Table 2: Accuracy (%) comparison of SOTA distillation methods on ImageNet.

Teacher	Student	Acc	Teacher	Teacher*	Student	KD	AT	CC	SP	RKD	CRD	SSKD	Ours	Ours*
ResNet-34	ResNet-18	Top-1	73.31	75.48	69.75	70.66	70.70	69.96	70.62	71.34	71.38	71.62	72.16	**72.39**
		Top-5	91.42	92.67	89.07	89.88	90.00	89.17	89.80	90.37	90.49	90.67	90.85	**91.00**

Table 3: Linear classification accuracy (%) of transfer learning.

Transferred Dataset	Baseline	KD	FitNet	AT	AB	VID	RKD	SP	CC	CRD	SSKD	Ours
CIFAR-100→ STL-10	67.76	67.90	69.41	67.37	67.82	69.29	69.74	68.96	69.13	70.09	71.03	**74.66**
CIFAR-100→ TinyImageNet	34.69	34.15	36.04	34.44	34.79	36.09	37.21	35.69	36.43	38.17	39.07	**42.57**