Hierarchical Self-supervised Augmented Knowledge Distillation

Chuanguang Yang[1,2] Zhulin An[1] Linhang Cai[1,2] Yongjun Xu[1]

[1]Institute of Computing Technology, Chinese Academy of Sciences

[2]University of Chinese Academy of Sciences

IJCAI, 19th-26th August 2021

Table of Contents

Table of Contents

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

Aspect 1

what kind of knowledge encapsulated in teacher network can be explored for KD:

1. **Logit-based class posterior distributions** (Hinton's KD [1])
2. **Feature-based information**: feature-maps (FitNet [2]), attention maps (AT [3]), gram matrix (FSP [4]), activation boundaries (AB [5]) and so on.
3. **Cross-sample relation information**: distance and angle relation (RKD [6]), correlation (CC [7]), similiarity (SP [1]), contrastive representations (CRD [2]), and self-supervised contrastive relations (SSKD [3]).

Introduction

Previous state-of-the-art: SSKD [3]

Inspired by SimCLR [4], SSKD [3] applies contrastive learning by forcing the image and its transformed version closed against other negative images. It defines the contrastive relationship as knowledge.

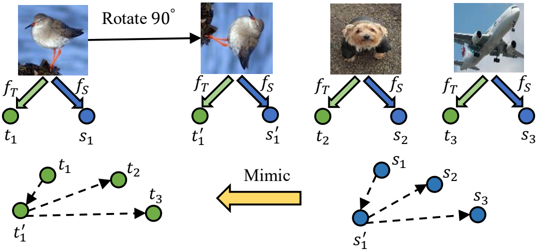


Figure: Self-supervised contrastive relationship [3].

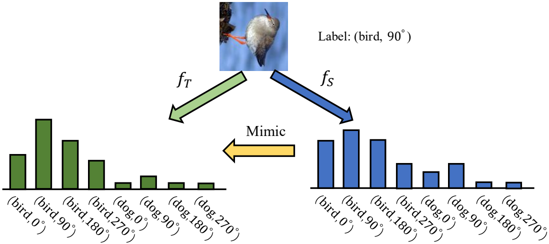
Introduction

Motivation

SSKD [3] learns invariant feature representations among transformed images using a self-supervised pretext task with random rotations from $0°, 90°, 180°, 270°$, which may destroy the original visual semantics.

A new type knowledge with self-supervised auxiliary tasks

We introduce a self-supervised augmented distribution that encapsulates the unified knowledge of the original classification task and auxiliary self-supervised task as the richer dark knowledge for the field of KD.



Empirical verification

We conduct initial exploratory experiments to train a ResNet-18 using rotation as a data augmentation (DA) and a self-supervised augmented label (SAL) as follows. The good performance by SAL further motivates us to define the self-supervised augmented distribution as a promising knowledge for KD.

Dataset	Baseline	+DA (Rotation)	+SAL (Rotation)
CIFAR-100	78.01	$77.75_{(\downarrow -0.26)}$	$\mathbf{79.76}_{(\uparrow +1.75)}$
TinyImageNet	63.69	$62.66_{(\downarrow -1.03)}$	$\mathbf{65.81}_{(\uparrow +2.12)}$

Aspect 2

How to effectively transfer knowledge from teacher to student:

1. **KD from logits from the final layer**: Hinton's KD [1]
2. **KD from intermediate feature-maps**: FitNet [2], AT [3], FSP [4] and AB [5].
3. **KD from highly abstract feature embeddings before the penultimate layers**: RKD [6], CC [7], SP [1], CRD [2] and SSKD [3].
4. **KD assisted with an extra teacher**: HKD [5] introduces an extra teacher model to further bridge the knowledge gap.

Introduction

Motivation

Compared with feature information, the probability distribution is indeed a more robust knowledge for KD [2]. However, it is difficult to explicitly derive probability distributions from hidden layers over the original architecture.

Our method

Therefore a natural idea is to append several auxiliary classifiers to the network at various hidden layers to generate multi-level probability distributions from hierarchical feature maps.

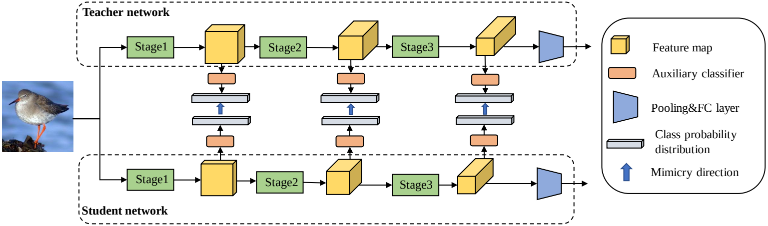


Table of Contents

CNN Architecture: A CNN can be decomposed into a feature extractor $\Phi(\cdot;\boldsymbol{\mu})$ and a linear classifier $g(\cdot;\boldsymbol{w})$, where $\boldsymbol{\mu}$ and $\boldsymbol{w}$ are weight tensors.

Conventional Class Probability Distribution

Given an input sample $\boldsymbol{x} \in \mathcal{X}$, $\mathcal{X}$ is the training set, $\boldsymbol{z} = \Phi(\boldsymbol{x};\boldsymbol{\mu}) \in \mathbb{R}^d$ is the extracted feature embedding vector, where d is the embedding size. We consider a conventional N-way object classification task with the label space $\mathcal{N} = \{1, \cdots, N\}$. The linear classifier maps the feature embedding $\boldsymbol{z}$ to a predictive class probability distribution $\boldsymbol{p}(\boldsymbol{x};\tau) = \sigma(g(\boldsymbol{z};\boldsymbol{w})/\tau) \in \mathbb{R}^N$.

Hinton's KD [1] uses conventional class proability distribution as knowledge.

Self-supervised Augmented Label Space

Assuming that we define M various image transformations $\{t_j\}_{j=1}^{M}$ with the label space $\mathcal{M} = \{1, \cdots, M\}$, where $t_1(\boldsymbol{x}) = \boldsymbol{x}$. The label space of this task is $\mathcal{K} = \mathcal{N} \otimes \mathcal{M}$, here $\otimes$ is the Cartesian product. $|\mathcal{K}| = N * M$, where $|\cdot|$ is the cardinality of the label collection, * denotes element-wise multiplication.

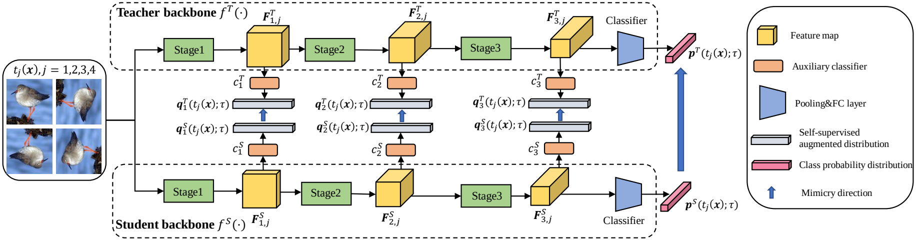
Self-supervised Augmented Distribution

Given a transformed sample $\tilde{\boldsymbol{x}} \in \{t_j(\boldsymbol{x})\}_{j=1}^{M}$ by applying one transformation on $\boldsymbol{x}$, $\tilde{\boldsymbol{z}} = \Phi(\tilde{\boldsymbol{x}}; \boldsymbol{\mu}) \in \mathbb{R}^d$ is the extracted feature embedding vector, $\boldsymbol{q}(\tilde{\boldsymbol{x}}; \tau) = \sigma(g(\tilde{\boldsymbol{z}}; \boldsymbol{w})/\tau) \in \mathbb{R}^{N*M}$ is the predictive distribution over the joint label space $\mathcal{K}$, where weight tensor $\boldsymbol{w} \in \mathbb{R}^{(N*M)\times d}$.

Methodology

Hierarchical Self-supervised Augmented Knowledge Distillation

We guide all auxiliary classifiers attached to the original network to learn informative self-supervised augmented distributions. Furthermore, we perform knowledge distillation between teacher and student towards all auxiliary classifiers in a one-to-one manner.



Methodology

Pre-train a teacher network

We denote the teacher network as $f^T(\cdot)$ and L auxiliary classifiers as $\{c_l^T(\cdot)\}_{l=1}^L$.

1. we train the $f^T(\cdot)$ with normal data $\boldsymbol{x}$ by the conventional Cross-Entropy (CE) loss to fit the ground-truth label $y \in \mathcal{N}$.
2. we aim to train L auxiliary classifiers $\{c_l^T(\cdot)\}_{l=1}^L$ for learning self-supervised augmented labels k_j.

The overall loss for training a teacher is shown in Eq. (1).

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

$p^T(x;\tau) = \sigma(f^T(x)/\tau) \in \mathbb{R}^N$ is predictive class probaility distribution, $q_l^T(t_j(x);\tau) = \sigma(c_l^T(F_{l,j}^T))/\tau) \in \mathbb{R}^{N*M}$ is self-supervised augmented distributions .

Methodology

Train a student network supervised by a teacher network

We denote the student backbone network as $f^S(\cdot)$ and L auxiliary classifiers as $\{c_l^S(\cdot)\}_{l=1}^L$. The overall loss includes a task loss from ground-truth labels and mimicry losses from the pre-trained teacher.

1. Task loss from ground-truth labels $\mathcal{L}_{task}$
2. Mimicry loss from self-supervised augmented distributions $\mathcal{L}_{kl\_q}$
3. Mimicry loss from class probability distributions $\mathcal{L}_{kl\_p}$

Methodology

Task loss from ground-truth labels $\mathcal{L}_{task}$

We force the $f^S(\cdot)$ to fit the normal data x with the ground-truth y as the task loss:

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

Where $p^S(x;\tau) = \sigma(f^S(x)/\tau) \in \mathbb{R}^N$ is the predictive class probaility distribution.

Methodology

Mimicry loss from self-supervised augmented distributions $\mathcal{L}_{kl\_q}$

We consider transferring hierarchical self-supervised augmented distributions $\{q_l^T(t_j(x);\tau)\}_{l=1}^{L}$ generated from L auxiliary classifiers of the teacher network to corresponding $\{q_l^S(t_j(x);\tau)\}_{l=1}^{L}$ generated from L auxiliary classifiers of the student network, respectively. The transfer performs in a one-to-one manner by KL-divergence loss D_{KL}.

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

Mimicry loss from class probability distributions $\mathcal{L}_{kl\_p}$

We transfer the original class probability distributions generated from the final layer between teacher and student. Specifically, we transfer the knowledge derived from both the normal and transformed data $\{t_j(x)\}_{j=1}^{M}$, where $t_1(x) = x$.

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

Methodology

Overall loss for training the student network

We summarize the task loss and mimicry loss as the overall loss $\mathcal{L}_S$ for training the student network:

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

Following the wide practice, we set the hyper-parameter $\tau = 1$ in task loss and $\tau = 3$ in mimicry loss. Besides, we do not introduce other hyper-parameters.

Table of Contents

Experiments

Experiments on CIFAR-100

Our HSAKD significantly outperforms the best-competing method SSKD across all network pairs with an average accuracy gain of 2.56%

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	$73.57_{(\pm0.23)}$	$71.95_{(\pm0.59)}$	$69.62_{(\pm0.26)}$	$72.95_{(\pm0.24)}$	$73.51_{(\pm0.26)}$	$73.51_{(\pm0.26)}$	$71.74_{(\pm0.35)}$	$72.96_{(\pm0.33)}$
KD	$75.23_{(\pm0.23)}$	$73.90_{(\pm0.44)}$	$70.91_{(\pm0.10)}$	$73.54_{(\pm0.26)}$	$75.21_{(\pm0.24)}$	$75.80_{(\pm0.46)}$	$75.83_{(\pm0.18)}$	$75.43_{(\pm0.33)}$
FitNet	$75.30_{(\pm0.42)}$	$74.30_{(\pm0.42)}$	$71.21_{(\pm0.16)}$	$75.37_{(\pm0.12)}$	$75.42_{(\pm0.34)}$	$75.41_{(\pm0.07)}$	$76.27_{(\pm0.18)}$	$76.91_{(\pm0.06)}$
AT	$75.64_{(\pm0.31)}$	$74.32_{(\pm0.23)}$	$71.35_{(\pm0.09)}$	$75.06_{(\pm0.19)}$	$74.08_{(\pm0.21)}$	$76.57_{(\pm0.20)}$	$76.51_{(\pm0.44)}$	$76.32_{(\pm0.12)}$
AB	$71.26_{(\pm1.32)}$	$74.55_{(\pm0.46)}$	$71.56_{(\pm0.19)}$	$74.31_{(\pm0.09)}$	$74.98_{(\pm0.44)}$	$75.87_{(\pm0.39)}$	$76.43_{(\pm0.09)}$	$76.40_{(\pm0.29)}$
VID	$75.31_{(\pm0.22)}$	$74.23_{(\pm0.28)}$	$71.35_{(\pm0.09)}$	$75.07_{(\pm0.35)}$	$75.67_{(\pm0.13)}$	$75.97_{(\pm0.08)}$	$76.24_{(\pm0.44)}$	$75.98_{(\pm0.41)}$
RKD	$75.33_{(\pm0.14)}$	$73.90_{(\pm0.26)}$	$71.67_{(\pm0.08)}$	$74.17_{(\pm0.22)}$	$75.54_{(\pm0.36)}$	$76.20_{(\pm0.06)}$	$75.74_{(\pm0.32)}$	$75.42_{(\pm0.25)}$
SP	$74.35_{(\pm0.59)}$	$72.91_{(\pm0.24)}$	$71.45_{(\pm0.38)}$	$75.44_{(\pm0.11)}$	$75.68_{(\pm0.35)}$	$76.35_{(\pm0.14)}$	$76.40_{(\pm0.37)}$	$76.43_{(\pm0.21)}$
CC	$75.30_{(\pm0.03)}$	$74.46_{(\pm0.05)}$	$71.44_{(\pm0.10)}$	$74.40_{(\pm0.24)}$	$75.66_{(\pm0.33)}$	$76.05_{(\pm0.25)}$	$75.63_{(\pm0.30)}$	$75.74_{(\pm0.18)}$
CRD	$75.81_{(\pm0.33)}$	$74.76_{(\pm0.25)}$	$71.83_{(\pm0.42)}$	$75.77_{(\pm0.24)}$	$76.13_{(\pm0.16)}$	$76.89_{(\pm0.27)}$	$76.37_{(\pm0.23)}$	$76.51_{(\pm0.09)}$
SSKD	$76.16_{(\pm0.17)}$	$75.84_{(\pm0.04)}$	$70.80_{(\pm0.02)}$	$75.83_{(\pm0.29)}$	$76.21_{(\pm0.16)}$	$78.21_{(\pm0.16)}$	$76.71_{(\pm0.31)}$	$77.64_{(\pm0.24)}$
Ours	$77.20_{(\pm0.17)}$	$77.00_{(\pm0.21)}$	$72.58_{(\pm0.33)}$	$77.26_{(\pm0.14)}$	$77.45_{(\pm0.21)}$	$78.79_{(\pm0.11)}$	$78.51_{(\pm0.20)}$	$79.93_{(\pm0.11)}$
Ours*	$\mathbf{78.67}_{(\pm0.20)}$	$\mathbf{78.12}_{(\pm0.25)}$	$\mathbf{73.73}_{(\pm0.10)}$	$\mathbf{77.69}_{(\pm0.05)}$	$\mathbf{79.27}_{(\pm0.12)}$	$\mathbf{79.43}_{(\pm0.24)}$	$\mathbf{80.11}_{(\pm0.32)}$	$\mathbf{80.86}_{(\pm0.15)}$

Table: Top-1 accuracy (%) comparison of SOTA methods.

Experiments

Experiments on ImageNet

Our HSAKD significantly outperforms the best-competing method SSKD on the pair of ResNet-34 and ResNet-18 with a top-1 accuracy gain of 0.77%

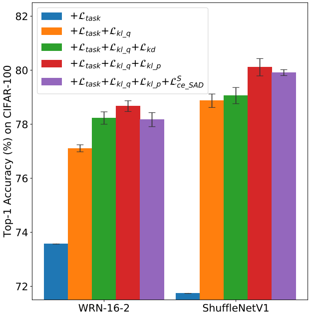
Teacher	Student	Acc	Teacher	Teacher*	Student	KD	AT	CC	SP	RKD	CRD	SSKD	Ours	Ours*
ResNet-34	ResNet-18	Top-1	73.31	75.48	69.75	70.66	70.70	69.96	70.62	71.34	71.38	71.62	72.16	**72.39**
		Top-5	91.42	92.67	89.07	89.88	90.00	89.17	89.80	90.37	90.49	90.67	90.85	**91.00**

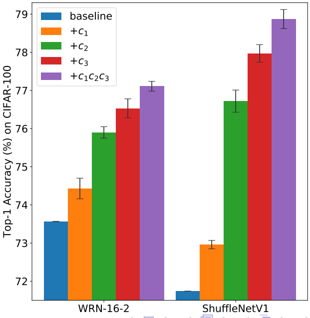
Table: Top-1 accuracy (%) comparison on ImageNet.

Experiments

Ablation study on CIFAR-100

- **Effect of loss terms *(left)*:** each loss term is indispensable, and $\mathcal{L}_{kl\_q}$ contributes most by distilling the self-supervised augmented distributions.
- **Effect 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.





Transfer Experiments on STL-10 and Tiny ImageNet

Our HSAKD can significantly outperform the best-competing SSKD by 3.63% on STL-10 and 3.50% on TinyImageNet.

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**

Table: Linear classification accuracy (%) of transfer learning on the student MobileNetV2 pre-trained using the teacher VGG-13.

Transfer Experiments on Pascal VOC

Our HSAKD outperforms the original baseline by 2.27% mAP and the best-competeting SSKD by 0.85% mAP on downstream object detection.

Baseline	KD	CRD	SSKD	Ours
76.18	77.06	77.36	<u>77.60</u>	**78.45**

Table: Comparison of detection mAP (%) on Pascal VOC using ResNet-18 as the backbone pre-trained by various KD methods.

Efficacy under Few-shot Scenario

Our HSAKD can consistently surpass other KD methods by large margins under various few-shot settings.

Percentage	KD	CRD	SSKD	Ours
25%	65.15$_{(\pm 0.23)}$	65.80$_{(\pm 0.61)}$	<u>67.82</u>$_{(\pm 0.30)}$	**68.50**$_{(\pm 0.24)}$
50%	68.61$_{(\pm 0.22)}$	69.91$_{(\pm 0.20)}$	<u>70.08</u>$_{(\pm 0.13)}$	**72.18**$_{(\pm 0.41)}$
75%	70.34$_{(\pm 0.09)}$	<u>70.98</u>$_{(\pm 0.43)}$	70.47$_{(\pm 0.14)}$	**73.26**$_{(\pm 0.11)}$

Table: Top-1 accuracy (%) comparison on CIFAR-100 under few-shot scenario with various percentages of training samples. We use the ResNet56-ResNet20 as the teacher-student pair for evaluation.

中科院计算所 INSTITUTE OF COMPUTING TECHNOLOGY,CAS

Table of Contents

Conclusion

1. **Knowledge defination**: We introduce a self-supervised augmented distribution that encapsulates the unified knowledge of the original classification task and auxiliary self-supervised task as the richer dark knowledge for the field of KD.
2. **Knowledge transfer**: 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.
3. **Experimental results**: 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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean.
Distilling the knowledge in a neural network.
arXiv preprint arXiv:1503.02531, 2015.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio.
Fitnets: Hints for thin deep nets.
ICLR, 2015.

Sergey Zagoruyko and Nikos Komodakis.
Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer.
ICLR, 2017.

Junho Yim, Donggyu Joo, Ji-Hoon Bae, and Junmo Kim.
A gift from knowledge distillation: Fast optimization, network minimization and transfer learning.
In *CVPR*, pages 7130–7138, 2017.

Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi.
Knowledge transfer via distillation of activation boundaries formed by hidden neurons.
In *AAAI*, volume 33, pages 3779–3787, 2019.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho.
Relational knowledge distillation.
In *CVPR*, pages 3967–3976, 2019.

Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang.
Correlation congruence for knowledge distillation.
In *ICCV*, pages 5007–5016, 2019.

Reference

Frederick Tung and Greg Mori.
Similarity-preserving knowledge distillation.
In *ICCV*, pages 1365–1374, 2019.

Yonglong Tian, Dilip Krishnan, and Phillip Isola.
Contrastive representation distillation.
ICLR, 2020.

Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy.
Knowledge distillation meets self-supervision.
In *ECCV*, pages 588–604, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton.
A simple framework for contrastive learning of visual representations.
ICML, pages 1597–1607, 2020.

Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas.
Heterogeneous knowledge distillation using information flow modeling.
In *CVPR*, pages 2339–2348, 2020.

Gidaris, Spyros and Singh, Praveer and Komodakis, Nikos.
Unsupervised representation learning by predicting image rotations.
In *ICLR*, 2018