Distilling Knowledge from Self-Supervised Teacher by Embedding Graph Alignment

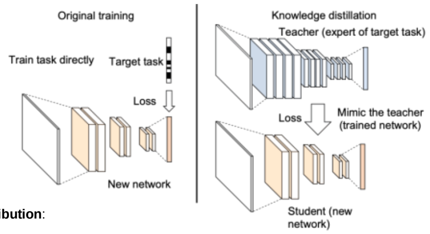
Yuchen Ma*[1], Yanbei Chen*[1], Zeynep Akata[1,2,3]

[1]University of Tübingen, [2]MPI for Informatics, [3]MPI for Intelligent Systems

*Equal contribution. The work was done when Yuchen Ma was enrolled in the MSc program at the University of Heidelberg.

Introduction

Task Introduction Knowledge Distillation

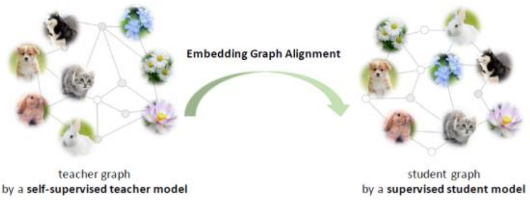


Contribution:

- Propose a new knowledge distillation method to transfer the instance-wise structural knowledge.
- Establish a comprehensive benchmark on three image classification datasets
- Demonstrate the superiority of our model under a variety of evaluation setups.

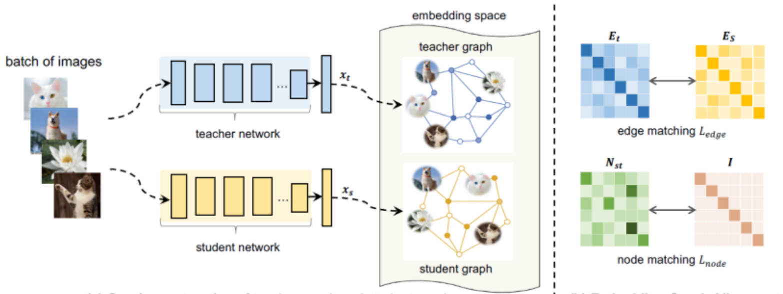
Motivation

Goal: Learn visual representation by knowledge distillation



- Modeling the **instance-instance correlations**
- Transferring the **graph structural knowledge**
- Use self-supervised knowledge

Model Overview



(a) **Graph construction** of teacher graph and student graph

(b) **Embedding Graph Alignment**

- **Construct** the teacher graph and the student graph
- **Align** the teacher graph and the student graph
- **Jointly optimize** an edge matching constraint and a node matching constraint.

Graph Construction

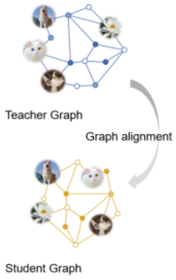
- **Node**:
Feed the extracted features to individual **node embedding layers**.

- **Edge**:
based on the Pearson's correlation coefficient (PPC)

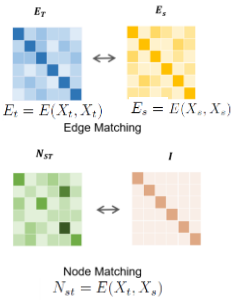
$$e_{x,y} = \frac{\sum_{i=1}^{D}(\mathrm{x}_i - \bar{\mathrm{x}})(\mathrm{y}_i - \bar{\mathrm{y}})}{\sqrt{\sum_{i=1}^{D}(\mathrm{x}_i - \bar{\mathrm{x}})^2}\sqrt{\sum_{i=1}^{D}(\mathrm{y}_i - \bar{\mathrm{y}})^2}}$$

- **Edge matrix:**
encode the correlation between every pair of images among the same batch

$$E(X,X) = \begin{pmatrix} 1 & \cdots & e_{1,B} \\ \vdots & \ddots \quad e_{i,j} & \vdots \\ \vdots \quad e_{j,i} & \ddots & \vdots \\ e_{B,1} & \cdots \quad \cdots & 1 \end{pmatrix}$$



Embedding Graph Alignment



- **Edge matching loss**

$$\mathcal{L}_{edge} \triangleq \|E_t - E_s\|_2$$

- **Node matching loss**

$$\mathcal{L}_{node} \triangleq \|N_{st} - \mathcal{I}\|_2$$

- **Distillation loss**

$$\mathcal{L}_{EGA} = \mathcal{L}_{node} + \lambda \mathcal{L}_{edge}$$

- **Training loss**

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_{EGA}\mathcal{L}_{EGA}$$

Experiment

- **Evaluation on different network architectures**

Method	Same student different teacher			Same teacher different student		
	ViT-B/32	ViT-B/16	RN101	Resnet8x4	ShuffleNetV1	VGG13
KD[16]	71.55	71.99	64.77	71.55	72.90	75.20
FitNet[30]	73.93	74.13	74.14	73.93	nan	75.56
PKT [27]	73.86	73.55	72.21	73.86	75.31	75.55
RKD[26]	73.34	73.42	73.7	73.34	73.93	76.41
NCE [10]	74.30	74.41	73.69	74.30	73.99	76.42
IRG [20]	75.11	74.72	74.17	75.11	74.79	75.98
CRD[32]	75.73	75.68	75.13	75.73	75.54	76.83
CCL[10]	75.91	76.13	75.08	75.91	76.14	**77.68**
EGA	**76.65**	**76.30**	**75.41**	**76.65**	**76.24**	77.59

The teacher and student are trained simultaneously

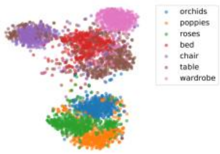
Method	Same student different teacher			Same teacher different student		
	ViT-B/32	ViT-B/16	RN101	Resnet8x4	ShuffleNetV1	VGG13
RKD[26]	73.36	72.43	73.92	73.36	72.62	73.26
CRD[32]	75.51	73.38	74.85	75.51	74.87	77.41
CCL[10]	75.98	39.56	74.22	75.98	76.05	77.54
EGA	**76.11**	**74.02**	**75.22**	**76.11**	**76.74**	**77.76**

The teacher and student are trained sequentially

- **Evaluation on supervised model**

Method	Same student different teacher		
	RN101	RN50	WRN-40
KD[16]	74.69	74.82	74.77
FitNet[30]	58.73	76.27	75.58
PKT[27]	74.44	75.69	75.30
RKD [26]	72.45	72.25	72.48
NCE[10]	73.62	74.35	72.90
CRD [32]	75.52	75.50	75.84
CCL[10]	75.56	75.53	75.33
EGA	**75.77**	**76.36**	**75.97**

- **Visualizing embeddings with t-SNE**

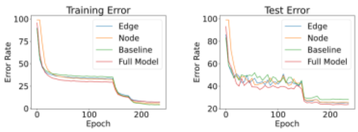


- **Evaluation on different dataset**

Method	CIFAR100	STL-10	TinyImageNet
KD [16]	71.55	84.35	54.68
FitNet [30]	76.04	84.15	59.97
PKT [27]	72.51	82.37	58.34
RKD [26]	73.34	83.13	58.15
NCE [10]	74.30	83.96	58.93
CRD [32]	75.73	82.40	60.34
CCL [10]	75.91	84.01	60.84
EGA	**76.65**	84.15	60.61
EGA + KD	76.49	**84.36**	**61.24**

Method	CIFAR 100	STL-10	TinyImageNet
RKD [26]	73.36	82.67	58.32
CRD [32]	75.51	78.76	60.82
CCL[10]	75.98	80.41	61.24
EGA	**76.11**	**83.01**	**61.85**

- **Analyzing the learning dynamics**



Reference

[10] Yanbei Chen, Yongqin Xian, A Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *CVPR*, 2021.
[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531, 2015.*
[26] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation.In *CVPR*, 2019.
[27] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *ECCV*, 2018.
[30] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015.
[32] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2019.