International Conference on Learning Representations-2025

Single Teacher, Multiple Perspectives: Teacher Knowledge Augmentation for Enhanced Knowledge Distillation

*Md Imtiaz Hossain, Sharmen Akhter, Choong Seon Hong * & Eui-Nam Huh **
Department of Computer Science & Engineering, Kyung Hee University, South Korea
Email: {hossain.imtiaz, sharmen, cshong, johnhuh}@khu.ac.kr

The Thirteenth International Conference on Learning Representations
Singapore - 2025

Paper: https://openreview.net/forum?id=DmEHmZ89iB

Code: https://github.com/mdimtiazh/TeKAP

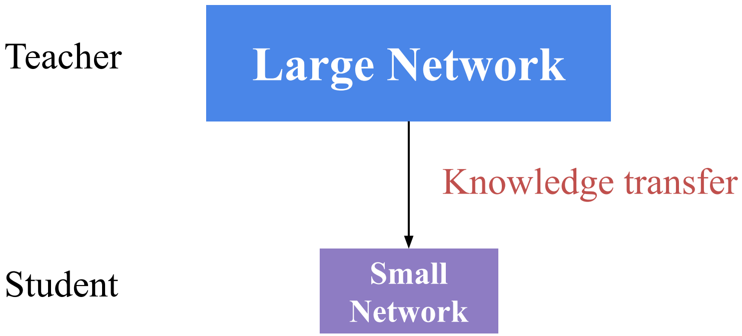
Contents

ICLR International Conference on Learning Representations-2025
ICNS Lab Intelligent-distributed Cloud and Security Lab
KYUNG HEE UNIVERSITY

Introduction

What is knowledge distillation

- Transfers knowledge from large to small model
- Model compression technique

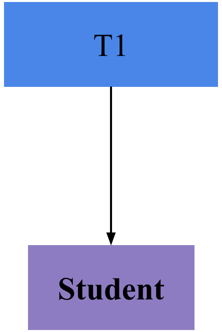


Single Teacher KD

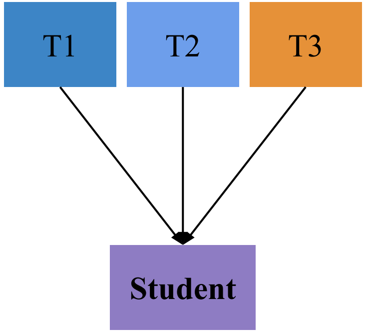
Motivation

Diversity & complexity analysis

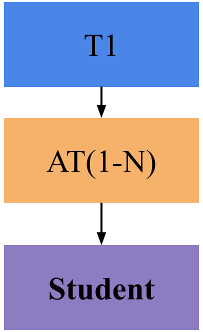
- Single teacher vs Multi-teacher



Single Teacher KD



Multi-teacher KD



TA based KD

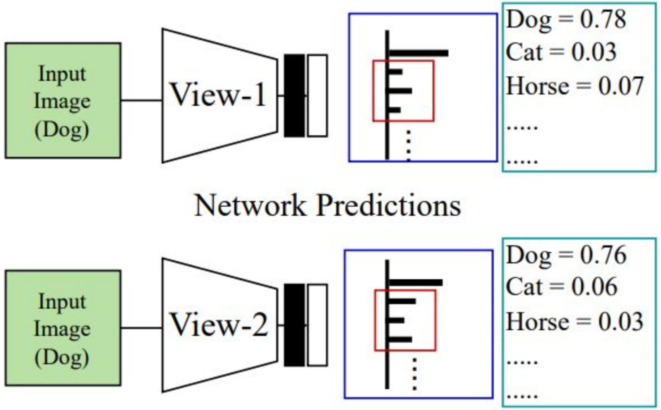
- Requires training single teachers
- Computationally less expensive
- Lacks diversity

- Enhances the performance of student
- Requires training multiple teachers
- Computationally expensive
- Provides multiple diversity

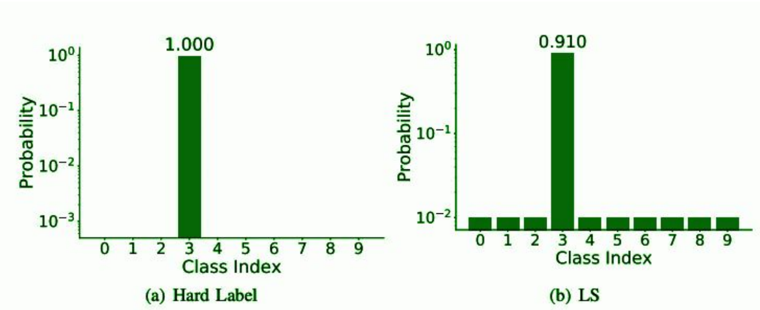
Problem Statement

One-hot vs label smoothing vs logits

- One-hot: impossible, rigid, no inter class relationship, no diversity
- Label smoothing: no inter class relationship, no diversity
- Teacher logit: inter-class relationship but lacks perspectives or diversity



Example of shifted inter-class relationship

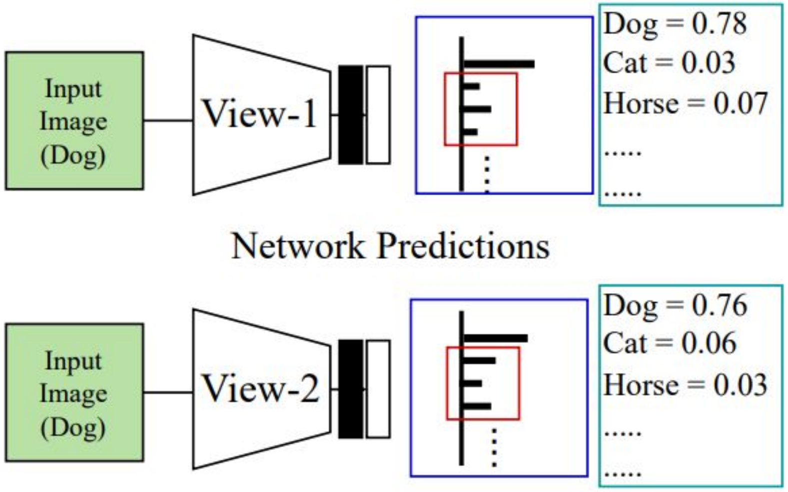


Understanding knowledge

Towards TeKAP

What multi-teacher based KD does?

- Differences in predictions
- Differences in inter-class relationships i.e., probabilities
- Differences in feature-level knowledge

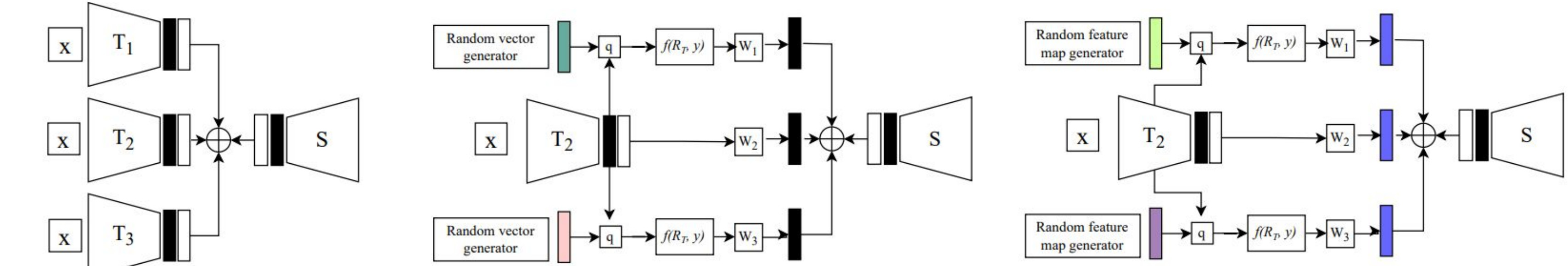


Example of shifted inter-class relationship

Generating multiple synthetic knowledge from one original teacher

Proposed Method

International Conference on Learning Representations-2025



(a) Traditional Multi-Teacher Distillation Approach

(b) Our Proposed Logit-Level TeKAP Approach

(c) Our Proposed Feature-Level TeKAP Approach

| S: Student Network | T: Teacher Network | x: Input | W: Linear weights | q: Linear weights for teacher and random vector |

TeKAP for (b) logit-level and (c) feature-level

Proposed Method

Feature-Level Distortion

$$f_T^{(i)}(x) = \alpha \times \eta_i + (1-\alpha) \times f_T(x)$$

Logit-Level Distortion

$$z_T^{(i)}(x) = \alpha \times \eta_i + (1-\alpha) \times z_T(x)$$

Total Feature-Level Loss

$$\mathcal{L}_{feat} = \lambda \mathrm{L}(f_S(x), f_T(x)) + (1-\lambda)\sum_{i=1}^{N} \mathrm{L}(f_S(x), f_T^i(x))$$

Total Distillation Loss

$$\mathcal{L}_{TeKAP} = \alpha \mathcal{L}_{feat} + \beta \mathcal{L}_{logit} + \gamma \mathcal{L}_{cel}$$

Total Logit-Level Loss

$$\mathcal{L}_{logits}^{\text{perturb}} = \lambda \mathcal{L}_{KD}(z_S(x), z_T(x)) + (1-\lambda)\sum_{i=1}^{N} \mathcal{L}_{KD}(z_S(x), z_T^{(i)}(x))$$

Here, α, β, and γ are the balancing weights

Theoretical Justification

- Justification for Feature-level Perturbation
  - As a form of regularization [2]
  - Exposes student to a range of variations
  - Forced to learn a robust inductive bias
  - Mapping without being overconfident [3]

- Justification for Logit-level Perturbation
  - Noisy logits act as different perspectives
  - Different sets of combinations [1]
  - Generalize across multiple noisy versions
  - A broader range of decision boundaries

Experimental Results

Comparison with SOTA teaching assistant based approach

Teacher Student	resnet32x4 resnet8x4	WRN_40_2 WRN_40_1	WRN_40_2 WRN_16_2	VGG13 VGG8	resnet56 resnet20	resnet32x4 ShuffleNetV1	resnet32x4 ShuffleNetV2	WRN-40-2 ShuffleNetV1
TAKD	73.81	73.78	75.12	73.23	70.83	74.53	74.82	75.34
TeKAP (L)	**74.79**	**73.80**	**75.21**	**74.00**	**71.32**	**74.92**	**75.43**	**76.75**
TeKAP (F+L)	75.98	74.41	76.20	74.42	71.92	75.60	77.38	76.59

Effect on adversarial robustness

	Teacher	Student	KD	**KD + TeKAP (L)**
top-1	29.88	21.44	21.12	**22.47**
top-5	51.43	44.47	44.04	**45.72**

Comparison with KD on ImageNet

Set	Teacher	Student	KD	KD + TeKAP (L)
Top-1	26.69	30.25	29.59	**29.33**
Top-5	8.58	10.93	10.30	**10.08**

TeKAP on class imbalance dataset

Methods	resnet32x4-resnet8x4	WRN_40_2-WRN_16_2	VGG13-VGG8
Baseline (KD)	41.71	52.08	47.52
+ TeKAP (Ours)	46.42	52.72	51.25

Transferability to different dataset

Set	Student	KD	KD + TeKAP (L)
CIFAR100-STL10	70.33	71.01	**72.94**
CIFAR100-TinyImageNet	34.82	35.53	**35.81**

Experimental Results

Comparison with SOTAs

Baselines	Teacher Student	resnet32x4 resnet8x4	WRN_40_2 WRN_40_1
	Teacher Student	79.42 72.50	75.61 71.98
Single Teacher	DKD **+ TeKAP**	76.32 **76.59**	74.81 **75.33**
	MLKD **+ TeKAP**	77.08 **77.36**	75.35 **75.67**
Multi- Teacher	TAKD **+ TeKAP**	73.93 **74.81**	73.83 **74.37**
	CA-MKD **+ TeKAP**	75.90 **76.34**	74.56 **74.98**
	DGKD **+ TeKAP**	75.31 **76.17**	74.23 **75.14**

Effect of TeKAP on diverse network at logit and feature level

	To Similar Architecture					To Different Architecture		
Teacher Student	resnet32x4 resnet8x4	WRN_40_2 WRN_40_1	WRN_40_2 WRN_16_2	VGG13 VGG8	resnet56 resnet20	resnet32x4 ShuffleNetV1	resnet32x4 ShuffleNetV2	WRN-40-2 ShuffleNetV1
Teacher Student	79.42 72.50	75.61 71.98	75.61 73.26	74.64 70.36	72.34 69.06	79.42 70.50	74.64 70.36	75.61 70.50
KD **+ TeKAP (L)**	73.33 **74.79**	73.69 **73.80**	74.92 **75.21**	72.98 **74.00**	70.66 **71.32**	74.07 **74.92**	72.98 **75.43**	74.83 **76.75**
CRD **+ TeKAP (F)**	75.51 **75.65**	74.14 **74.21**	75.48 **75.83**	73.94 **74.10**	71.16 **71.71**	75.11 **75.55**	75.65 **76.23**	76.05 **76.60**
TeKAP (F+L)	75.98	74.41	76.20	74.42	71.92	75.60	77.38	76.59

Conclusion

- Improved knowledge sources
- Enhances diversity
- Explains why knowledge distillation works and leveraged controlled randomness
- Generates multiple synthetic teacher knowledge perspectives,
- Single teacher, multiple perspectives

- *Limitations & Future Works:*
  - Does not optimize the noise
  - Remains train-free
  - Plan to explore optimization-based techniques

References

1. Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. *Learning from noisy labels with deep neural networks: A survey.* IEEE transactions on neural networks and learning systems, 34 (11):8135–8153, 2022.
2. Yehui Tang, Yunhe Wang, Yixing Xu, Boxin Shi, Chao Xu, Chunjing Xu, and Chang Xu. *Beyond dropout: Feature map distortion to regularize deep neural networks.* In Proceedings of the AAAI conference on artificial intelligence, volume 34, pp. 5964–5971, 2020.
3. Zeyuan Allen-Zhu and Yuanzhi Li. *Towards understanding ensemble, knowledge distillation and self-distillation in deep learning.* arXiv preprint arXiv:2012.09816, 2020.
4. Rafael Muller, Simon Kornblith, and Geoffrey E Hinton. *When does label smoothing help?* ¨ Advances in neural information processing systems, 32, 2019.

Thank You!

The Thirteenth International Conference on Learning Representations
Singapore - 2025

Paper: https://openreview.net/forum?id=DmEHmZ89iB

Code: https://github.com/mdimtiazh/TeKAP