Knowledge Distillation from A Stronger Teacher

Tao Huang[1,2], Shan You[1], Fei Wang[1], Chen Qian[1], Chang Xu[2]

[1]SenseTime Research, [2]The University of Sydney

ArXiv preprint arXiv:2205.14589 (2022).

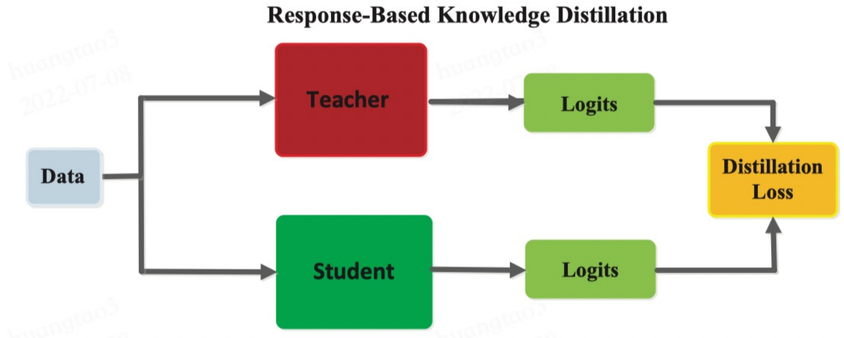
What is knowledge distillation?

Knowledge distillation (KD) is a model compression method in which a small model (student) is trained to distill knowledge from another model (teacher).

- KD was first proposed by[1] then generalized by[2].
- Generally, the teacher model is a pre-trained larger model.

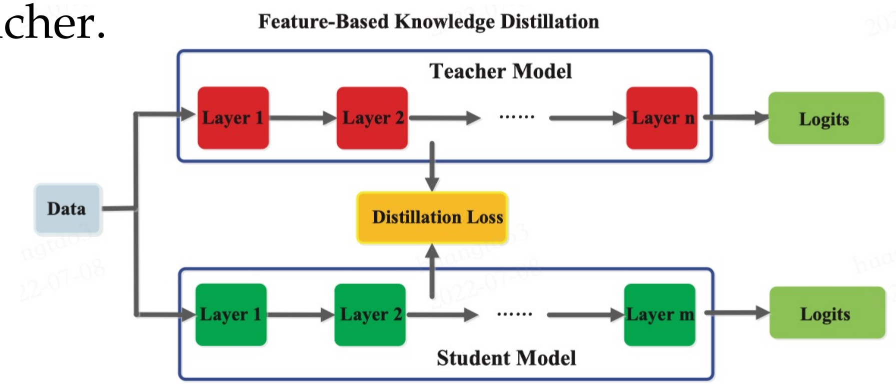
Response-based method

Distills knowledge in the outputs of the teacher.



Feature-based method

Distills knowledge in the intermediate features of the teacher.



[1]Buciluǎ, C., Caruana, R., & Niculescu-Mizil, A. (2006, August). Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 535-541).

[2]Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network.

Models are getting stronger

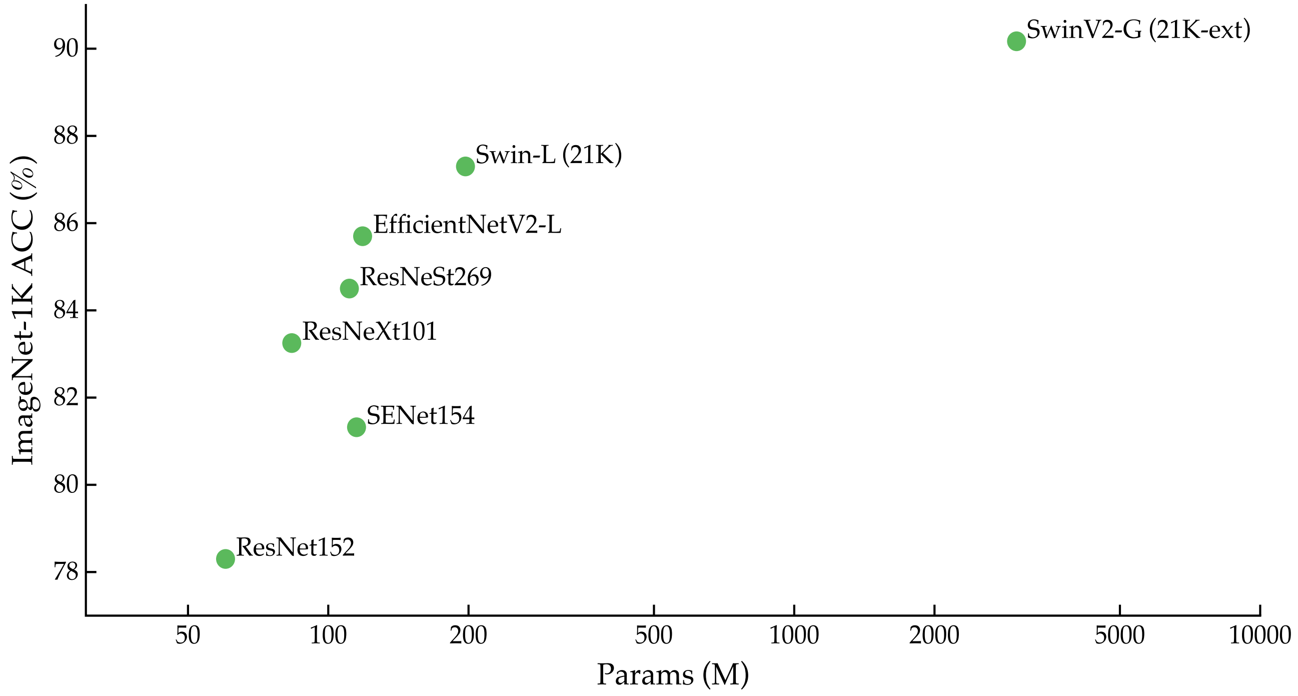
Evaluation settings of KD methods on ImageNet

Commonly-used settings:

- Models (teacher-student): ResNet34-ResNet18, ResNet50-MobileNetV1
- Training strategy: baseline (100 epochs, random crop, SGD optimizer, …)

Nevertheless, the ImageNet-1K performance has been greatly improved by designing larger models and stronger training strategies.

The baseline settings might be outdated and insufficient to today's practice.



Models are getting stronger

Evaluation settings of KD methods on ImageNet

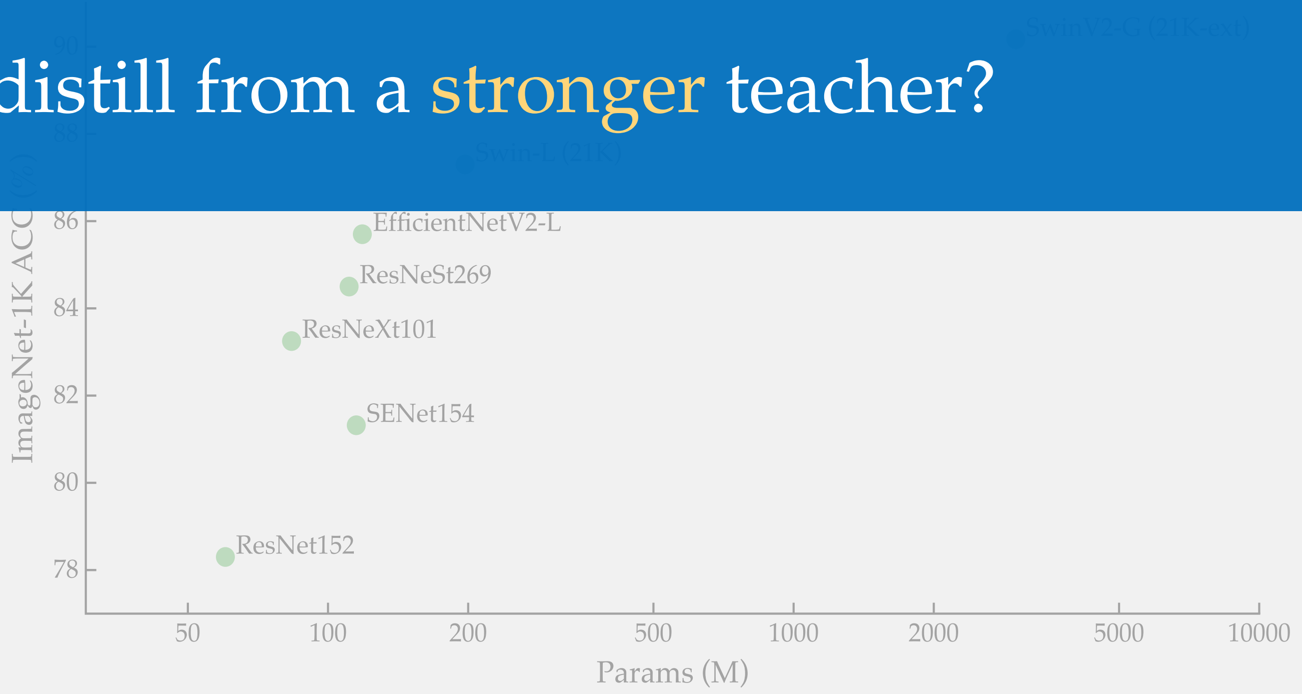
Commonly-used settings:

- Models (teacher-student): ResNet34-ResNet18, ResNet50-MobileNetV1
- Training strategy: baseline (100 epochs, random crop, SGD optimizer, …)

Would it be better to distill from a stronger teacher?

Nevertheless, the ImageNet-1K performance has been greatly improved by designing larger models and stronger training strategies.

The baseline settings might be outdated and insufficient to today’s practice.

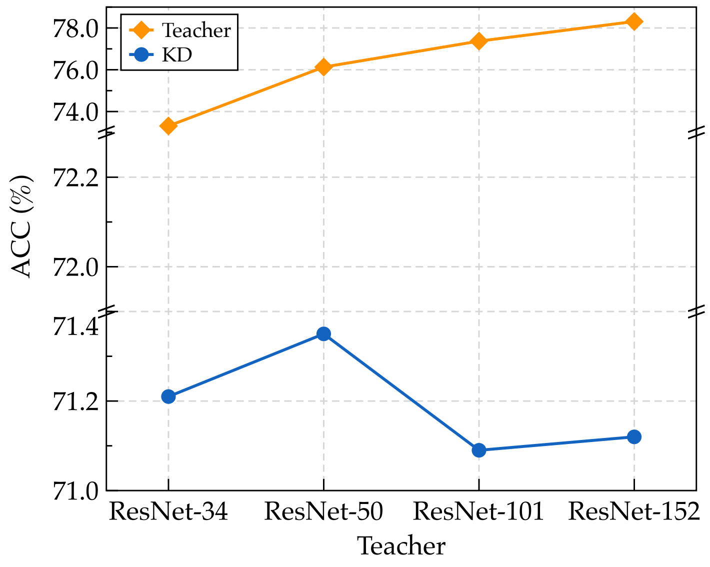


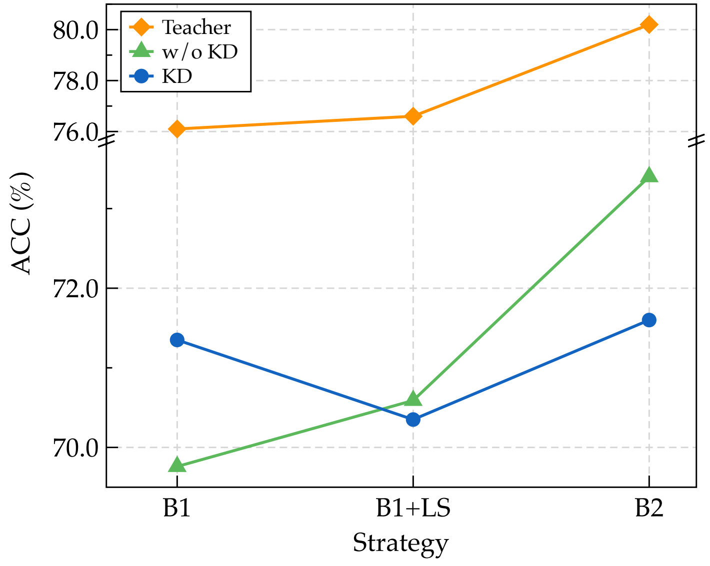
Unexpected performance drop with stronger teachers

Directly utilizing a stronger teacher in vanilla KD (KL div.):

Our experiments on ResNet-18 student and different teachers:

- Larger teachers: the ACCs of KD with R152 and R101 are lower than R34.
- Stronger strategies: the ACCs of KD with stronger strategies are even lower than standalone training.





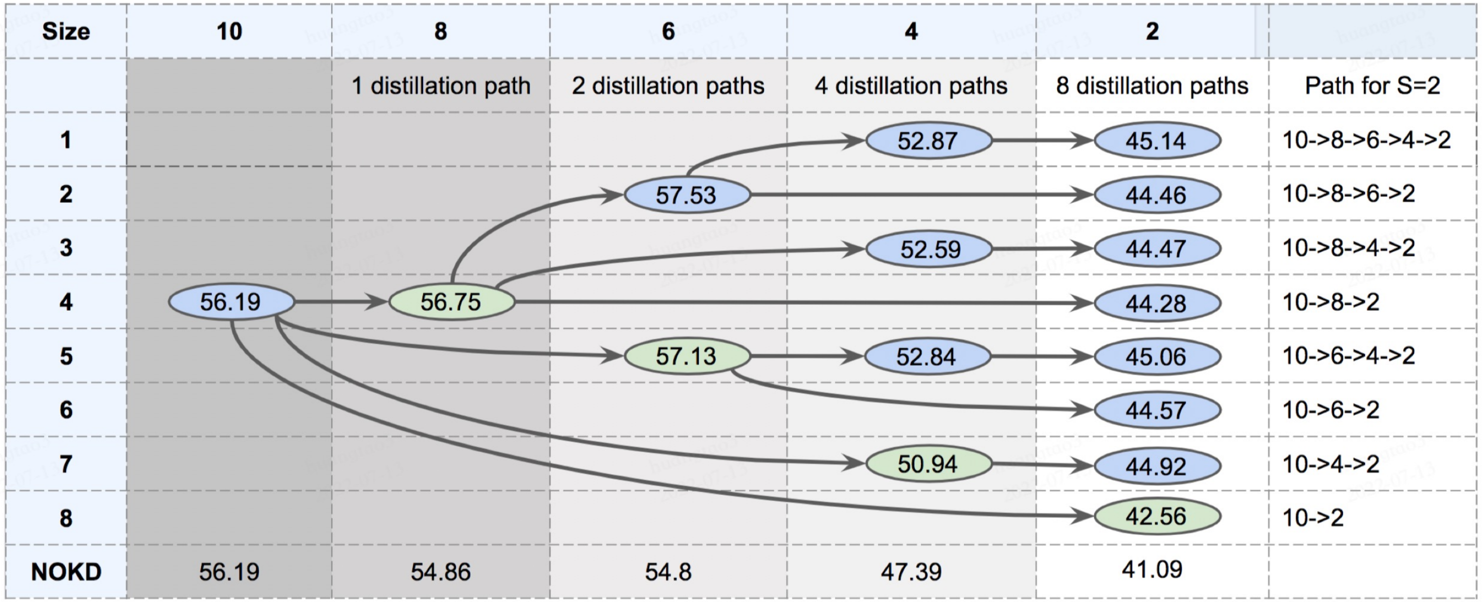
Conclusion:

- Stronger teachers ≠ better performance in vanilla KD.
- The effect of vanilla KD is severely affected by training strategy.

Teachers with larger capacities:

TAKD[3]: *a teacher can effectively transfer its knowledge to students up to a certain size.*

Solution: employ intermediate-sized networks as teacher assistants to bridge the gap between teacher and student.



Distillation paths for plain CNN on CIFAR-100

[3]Mirzadeh, S. I., Farajtabar, M., et al. (2020). Improved knowledge distillation via teacher assistant. *In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 04, pp. 5191-5198).*

What makes stronger teachers abnormal?

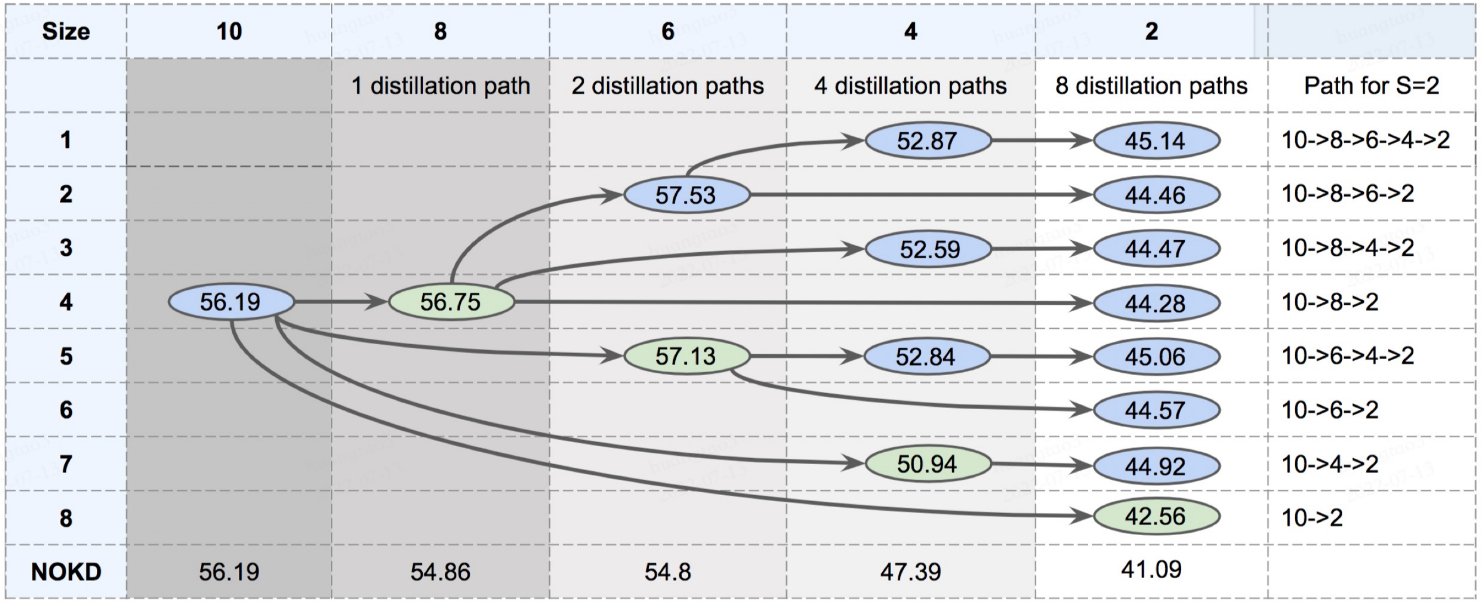
Teachers with larger capacities:

TAKD[3] : *a teacher can effectively transfer its knowledge to students up to a certain size.*

Solution: employ intermediate-sized networks as teacher assistants to bridge the gap between teacher and student.

Weaknesses:

- Need to train multiple models.
- The effect of KD is limited by the performance of teacher assistants.



Distillation paths for plain CNN on CIFAR-100

[3]Mirzadeh, S. I., Farajtabar, M., et al. (2020). Improved knowledge distillation via teacher assistant. *In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 04, pp. 5191-5198).*

What makes stronger teachers abnormal?

Teachers trained with stronger strategy:

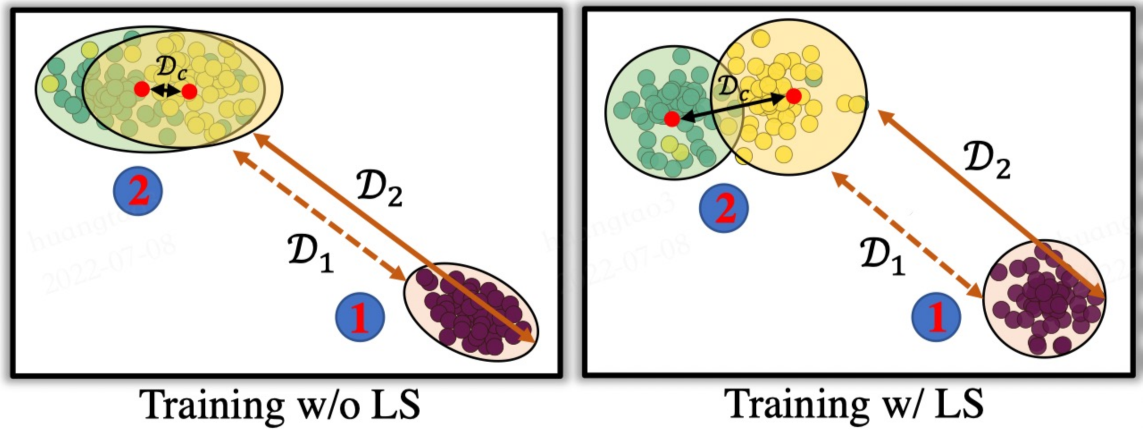
Previous works mainly focus on label smoothing (LS):

- Müller et al. (2019) [4]: *if a teacher network is trained with label smoothing, knowledge distillation into a student network is much less effective.*
- Shen et al. (2021) [5], Chandrasegaran, K., et al. (2022) [6]: *LS can be effective with KD (T=1).*

Observations of the effects of LS:

① LS enforces equidistant clusters (D_1 and D_2): weakening the relative information between logits.

② LS enlarges distances on those semantically similar classes.



T-SNE visualization of penultimate layer output.

[4]Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help?. *Advances in neural information processing systems, 32*.

[5]Shen, Z., Liu, Z., Xu, D., et al. (2021). Is Label Smoothing Truly Incompatible with Knowledge Distillation: An Empirical Study. *In International Conference on Learning Representations, 2021.*

[6]Chandrasegaran, K., et al. (2022). To Smooth or not to Smooth? On Compatibility between Label Smoothing and Knowledge Distillation. *https://openreview.net/forum?id=Vvmj4zGU_z3.*

Teachers trained with stronger strategy:

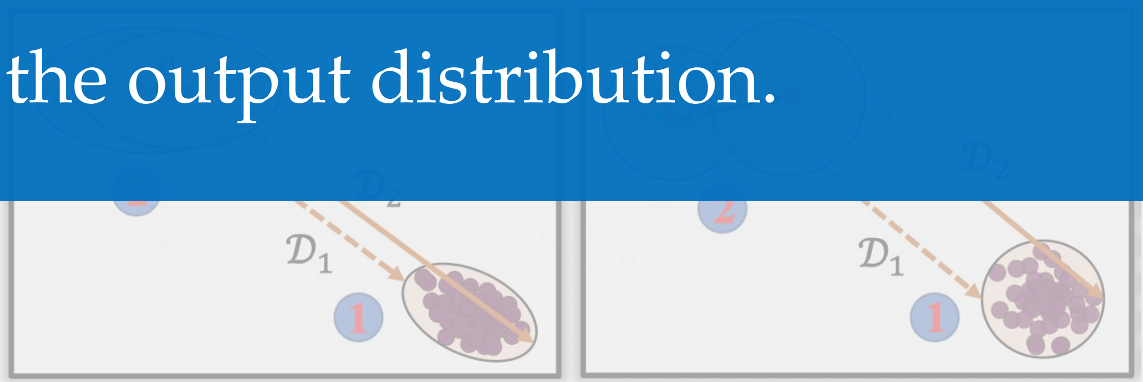
Previous works mainly focus on label smoothing (LS):

- Müller et al. (2019) [4]: *if a teacher network is trained with label smoothing, knowledge distillation into a student network is much less effective.*
- Shen et al. (2021) [5], Chandrasegaran, K., et al. (2022) [6]: *LS can be effective with KD (T=1).*

Label smoothing changes the output distribution.

Observations on the effects of LS

1. LS enforces equidistant clusters ($\mathcal{D}_1$ and $\mathcal{D}_2$): weakening the relative information between logits.
2. LS enlarges distances on those semantically similar classes.



Training w/o LS　　Training w/ LS

---

[4] Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help?. *Advances in neural information processing systems, 32*.

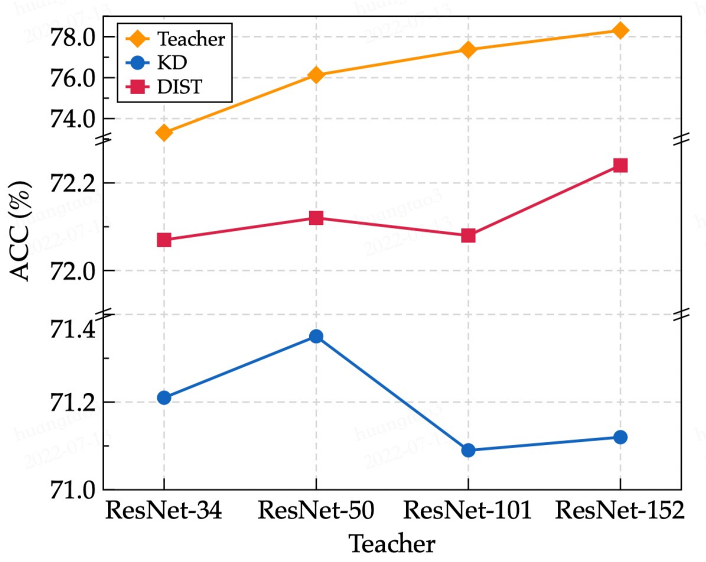
[5] Shen, Z., Liu, Z., Xu, D., et al. (2021). Is Label Smoothing Truly Incompatible with Knowledge Distillation: An Empirical Study. *In International Conference on Learning Representations, 2021.*

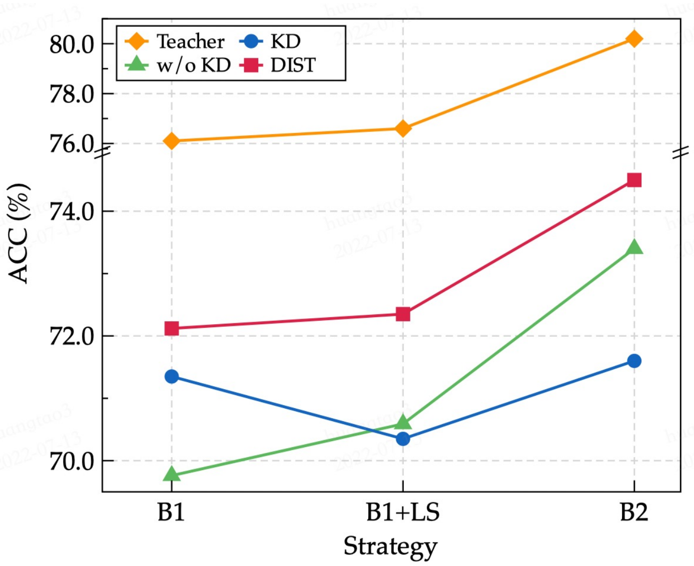
[6] Chandrasegaran, K., et al. (2022). To Smooth or not to Smooth? On Compatibility between Label Smoothing and Knowledge Distillation. *https://openreview.net/forum?id=Vvmj4zGU_z3.*

Contributions in our paper

In our paper (DIST):

- We unify teacher with larger capacity and teacher with stronger training strategy into one topic: stronger teacher, as they both **change the output distribution of teacher**.
- We extend the training strategies in KD with state-of-the-art strategies on CNNs and ViTs, *e.g.*, Label smoothing, AutoAugment, MixUp.
- We propose a new response-based KD method and show that, student's performance can be significantly boosted with a stronger teacher, without teacher assistants or sophisticated tuning on hyper-parameters (*e.g.*, temperature) in previous methods.



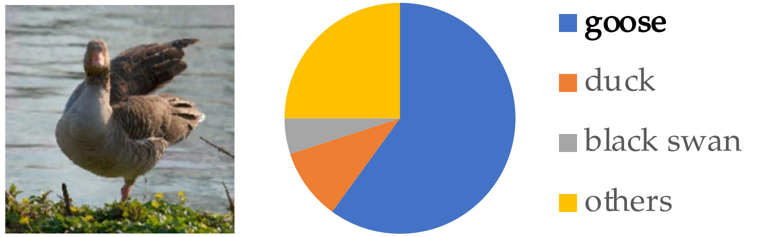


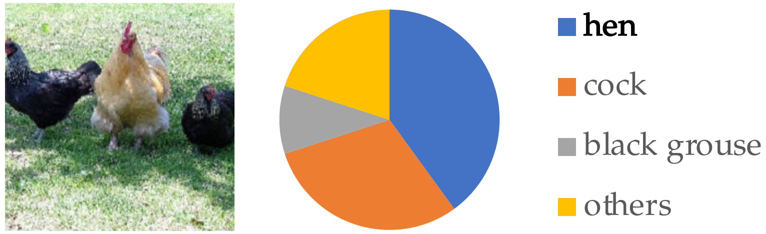
What do we truly care about for model's outputs?

In classification task, we care about:

- Which class has the largest probability for each sample.
- Fine-grained information: which classes are more related to the sample, etc.

We care more about relations rather than the exact values of outputs.





Kullback-Leibler (KL) divergence in KD:

$$\mathcal{L}_{\mathrm{KD}} := \frac{\tau^2}{B}\sum_{i=1}^{B}\mathrm{KL}(\boldsymbol{Y}^{(\mathrm{t})}_{i,:}, \boldsymbol{Y}^{(\mathrm{s})}_{i,:}) = \frac{\tau^2}{B}\sum_{i=1}^{B}\sum_{j=1}^{C} Y^{(\mathrm{t})}_{i,j}\log\left(\frac{Y^{(\mathrm{t})}_{i,j}}{Y^{(\mathrm{s})}_{i,j}}\right)$$

KL divergence matches the distribution point-wisely.

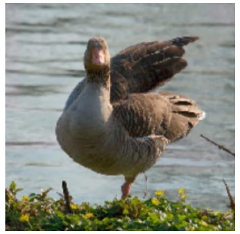
- It is vulnerable to the distribution changes.
- It conflicts with the Cross-Entropy loss of hard labels.

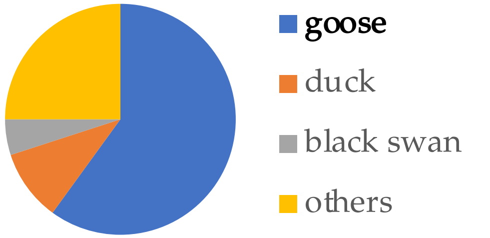
What do we truly care about for model’s outputs?

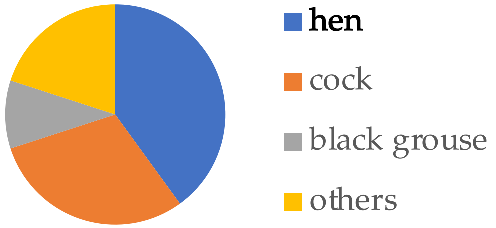
In classification task, we care about:

- Which class has the largest probability for each sample.
- Fine-grained information: which classes are more related to the sample, etc.

We care more about relations rather than the exact values of outputs.







Kullback-Leibler (KL) divergence in KD:

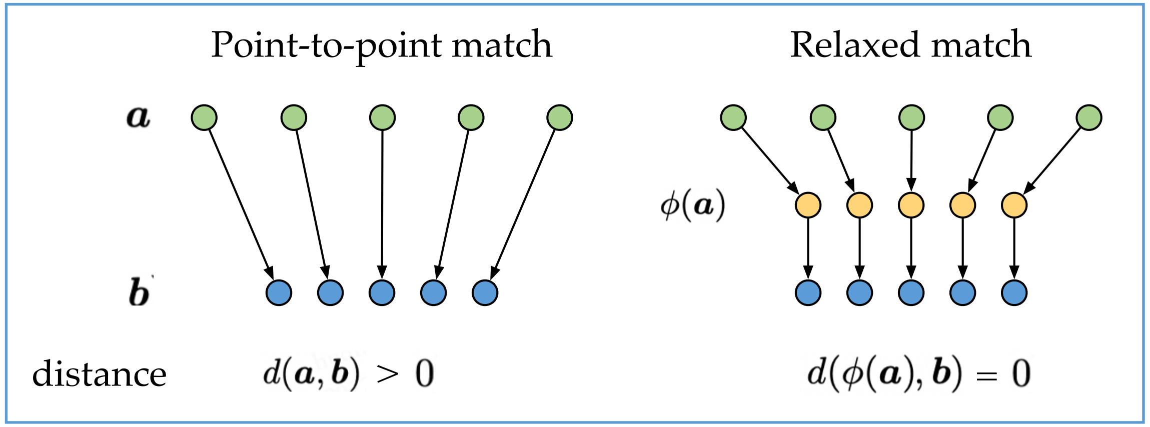
$$\mathcal{L}_{\mathrm{KD}} := \frac{\tau^2}{B}\sum_{i=1}^{B} \mathrm{KL}\left(\boldsymbol{Y}^{(\mathrm{t})}_{i,:}, \boldsymbol{Y}^{(\mathrm{s})}_{i,:}\right) = \frac{\tau^2}{B}\sum_{i=1}^{B}\sum_{j=1}^{C} Y^{(\mathrm{t})}_{i,j} \log\left(\frac{Y^{(\mathrm{t})}_{i,j}}{Y^{(\mathrm{s})}_{i,j}}\right)$$

KL divergence matches the distribution point-wisely.

- It is vulnerable to the distribution changes.
- It conflicts with the Cross-Entropy loss of hard labels.

We can just match the relations between teacher and student.

Relaxed match with relations



Considering that we have two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, and some distance metric $d(\cdot, \cdot)$ with $\mathbb{R}^C \times \mathbb{R}^C \to \mathbb{R}^+$ used to measure the discrepancy of $\boldsymbol{a}$ and $\boldsymbol{b}$.

For point-to-point matches such as KL divergence, $d(\boldsymbol{a}, \boldsymbol{b}) = 0$ if and only if $\boldsymbol{a} = \boldsymbol{b}$.

For a relaxed match, we want $d(\boldsymbol{a}, \boldsymbol{b}) = 0$ does not necessarily require $\boldsymbol{a}$ and $\boldsymbol{b}$ to be exactly the same.

Therefore, we can have additional mappings $\phi(\cdot)$ and $\psi(\cdot)$ with $\mathbb{R}^C \to \mathbb{R}^C$ such that

$$d(\phi(\boldsymbol{a}), \psi(\boldsymbol{b})) = d(\boldsymbol{a}, \boldsymbol{b}), \forall \boldsymbol{a}, \boldsymbol{b}$$

As a result, $d(\boldsymbol{a}, \boldsymbol{b})$ can be minimized when any of $d(\phi(\boldsymbol{a}), \psi(\boldsymbol{b}))$ gets minimized.

Relaxed match with relations

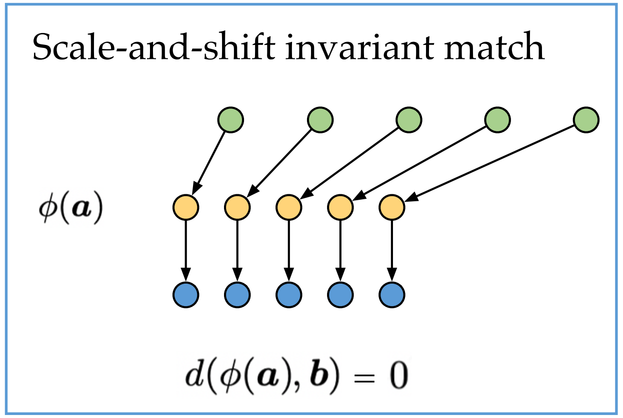
Pearson correlation for relative matching:

Since we care about the relation within $\boldsymbol{a}$ and $\boldsymbol{b}$, the mappings should be isotone and do not affect the semantic information and prediction results.

We choose a simple yet effective isotone mapping: linear transformation.
Therefore, the distance metric should satisfy

$$d(m_1\boldsymbol{a} + n_1, m_2\boldsymbol{b} + n_2) = d(\boldsymbol{a}, \boldsymbol{b}),$$

where $m_1, m_2, n_1,$ and n_2 are constants with $m_1 \times m_2 > 0$.

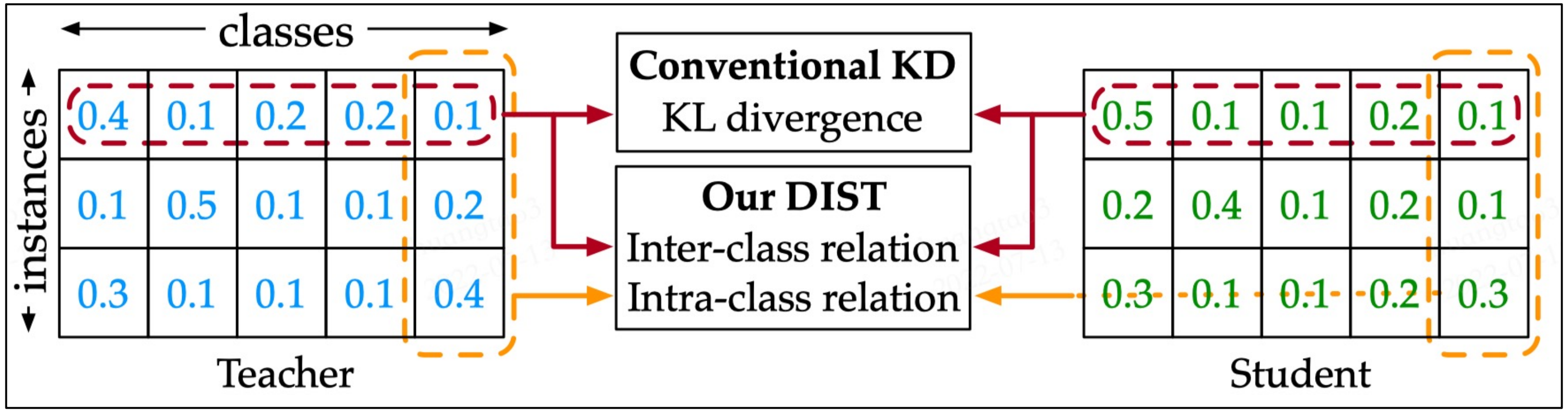


Pearson distance (centered cosine distance):

Pearson correlation coefficient is widely used to measure the linear correlation of two vectors, it is invariant under separate changes in location and scale in the two vectors.

$$d_{\mathrm{p}}(\boldsymbol{u}, \boldsymbol{v}) := 1 - \rho_{\mathrm{p}}(\boldsymbol{u}, \boldsymbol{v}) \quad \text{with} \quad \rho_{\mathrm{p}}(\boldsymbol{u}, \boldsymbol{v}) := \frac{\mathrm{Cov}(\boldsymbol{u}, \boldsymbol{v})}{\mathrm{Std}(\boldsymbol{u})\mathrm{Std}(\boldsymbol{v})} = \frac{\sum_{i=1}^{C}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{C}(u_i - \bar{u})^2 \sum_{i=1}^{C}(v_i - \bar{v})^2}}$$

Better distillation with inter-class and intra-class relations



By replacing the original KL divergence with Pearson distance, we have the following KD loss:

$$\mathcal{L}_{\mathrm{inter}} := \frac{1}{B}\sum_{i=1}^{B} d_{\mathrm{p}}(\mathbf{Y}_{i,:}^{(\mathrm{s})}, \mathbf{Y}_{i,:}^{(\mathrm{t})})$$

Considering that different samples have different similarities to each class, we further introduce a intra-class relation loss to transfer this relation.

$$\mathcal{L}_{\mathrm{intra}} := \frac{1}{C}\sum_{j=1}^{C} d_{\mathrm{p}}(\mathbf{Y}_{:,j}^{(\mathrm{s})}, \mathbf{Y}_{:,j}^{(\mathrm{t})})$$

Which one is more related to “cat”?

Overall training loss:

$$\mathcal{L}_{\mathrm{tr}} = \alpha\mathcal{L}_{\mathrm{cls}} + \beta\mathcal{L}_{\mathrm{inter}} + \gamma\mathcal{L}_{\mathrm{intra}}$$

Experimental setups

Table 1: **Training strategies on image classification tasks.** *BS*: batch size; *LR*: learning rate; *WD*: weight decay; *LS*: label smoothing; *EMA*: model exponential moving average; *RA*: RandAugment [8]; *RE*: random erasing; *CJ*: color jitter.

Strategy	Dataset	Epochs	Total BS	Initial LR	Optimizer	WD	LS	EMA	LR scheduler	Data augmentation
A1	CIFAR-100	240	64	0.05	SGD	5×10^{-4}	-	-	$\times 0.1$ at 150,180,210 epochs	crop + flip
B1	ImageNet	100	256	0.1	SGD	1×10^{-4}	-	-	$\times 0.1$ every 30 epochs	crop + flip
B2	ImageNet	450	768	0.048	RMSProp	1×10^{-5}	0.1	0.9999	$\times 0.97$ every 2.4 epochs	{*B1*} + RA + RE
B3	ImageNet	300	1024	5e-4	AdamW	5×10^{-2}	0.1	-	cosine	{*B2*} + CJ + Mixup + CutMix

We evaluate our DIST on various settings and tasks:

Image classification:
- CIFAR-100.
- Baseline settings on ImageNet.
- Larger teachers on ImageNet (ResNets).
- Stronger training strategies on ImageNet (ResNets, MobileNetV2, EfficientNet, Swin-Transformers).

Object detection

Semantic segmentation

Experiments on baseline settings

DIST significantly outperforms KD on baseline models and training strategies.

Table 2: **Evaluation results of baseline settings on ImageNet.** We use ResNet-34 and ResNet-50 released by Torchvision [27] as our teacher networks, and follow the standard training strategy (B1).

Student (teacher)		Teacher	Student	KD [15]	OFD [13]	CRD [40]	SRRL [46]	Review [7]	DIST
ResNet-18 (ResNet-34)	Top-1	73.31	69.76	70.66	71.08	71.17	71.73	71.61	**72.07**
	Top-5	91.42	89.08	89.88	90.07	90.13	90.60	90.51	90.42
MobileNet (ResNet-50)	Top-1	76.16	70.13	70.68	71.25	71.37	72.49	72.56	**73.24**
	Top-5	92.86	89.49	90.30	90.34	90.41	90.92	91.00	91.12

Training speed (batches/second):

KD [15]	RKD [29]	SRRL [46]	CRD [40]	DIST
14.28	11.11	12.98	8.33	14.19

Experiments on stronger teachers

Larger teachers:

Table 3: **Performance of ResNet-18 and ResNet-34 on ImageNet with different sizes of teachers.**

Student	Teacher	Top-1 ACC (%)			
		student	teacher	KD	DIST
ResNet-18	ResNet-34	69.76	73.31	71.21	**72.07** (+0.86)
	ResNet-50		76.13	71.35	**72.12** (+0.77)
	ResNet-101		77.37	71.09	**72.08** (+0.99)
	ResNet-152		78.31	71.12	**72.24** (+1.12)
ResNet-34	ResNet-50	73.31	76.13	74.73	**75.06** (+0.33)
	ResNet-101		77.37	74.89	**75.36** (+0.47)
	ResNet-152		78.31	74.87	**75.42** (+0.55)

Stronger training strategies:

Table 4: **Performance of students trained with strong strategies on ImageNet.** The *Swin-T* is trained with strategy B3 in Table 1, others are trained with B2. †: trained by [43]. ‡: Pretrained on ImageNet-22K.

Teacher	Student	Top-1 ACC (%)					
		teacher	student	KD [15]	RKD [29]	SRRL [46]	DIST
ResNet-50†	ResNet-18	80.1	73.4	72.6	72.9	71.2	**74.5**
	ResNet-34		76.8	77.2	76.6	76.7	**77.8**
	MobileNetV2		73.6	71.7	73.1	69.2	**74.4**
	EfficientNet-B0		78.0	77.4	77.5	77.3	**78.6**
Swin-L‡	ResNet-50	86.3	78.5	80.0	78.9	78.6	**80.2**
	Swin-T		81.3	81.5	81.2	81.5	**82.3**

Significant improvements on **small** models.

Ablation studies

Effects of inter-class and intra-class relations:

Method	Inter	Intra	ACC (%)
KD	-	-	71.21
DIST (KL div.)	✗	✓	70.61
DIST (KL div.)	✓	✓	71.62
DIST	✓	✗	71.63
DIST	✗	✓	71.55
DIST	✓	✓	**72.07**

Intra-class relation can also improve vanilla KD.

Training without task loss:

DIST is more informative than KD and ground-truth labels.

Method	w/ cls. loss	w/o cls. loss
KD	71.21	68.12
DIST	**72.07**	**70.65**

ResNet-18: 69.76%

Conclusion and future works

Conclusion:

We unify and analyze the performance collapse problem of stronger teachers in KD from a distribution match perspective.

We propose a new response-based KD method dubbed DIST to relax the distribution match, which

- adapts well on various models, strategies, tasks;
- is pretty simple and fast, and has the same training speed as KD;

Potential research directions:

- More stronger teachers: generic vision fundamental models.
- Better the relation mappings: rank correlations, non-linear mappings, etc.
- Training student-friendly teachers.
- …

Thank you!

Code is available at: https://github.com/hunto/DIST_KD

Questions: contact thua7590@uni.sydney.edu.au