Less or More From Teacher: Exploiting Trilateral Geometry For Knowledge Distillation

Chengming Hu[1,2*], Haolun Wu[1,2*], Xuan Li[1,2], Chen Ma[3], Xi Chen[1], Jun Yan[4], Boyu Wang[5], Xue Liu[1,2]

[1]McGill University, [2]Mila – Quebec AI Institute, [3]City University of Hong Kong, [4]Concordia University, [5]Western University

*Equal contribution with random order.

Background and Motivation

- **Knowledge distillation (KD)** [1] transfers knowledge from a large teacher to a lightweight student.

Objective: imitating the teacher's behaviors and matching the ground truths.

$$\mathcal{L} = \alpha \mathcal{L}^{\mathrm{KD}} + (1 - \alpha)\mathcal{L}^{\mathrm{GT}} \qquad \alpha \in [0, 1]$$

***Knowledge fusion ratio*:** the trade-off between two signals.

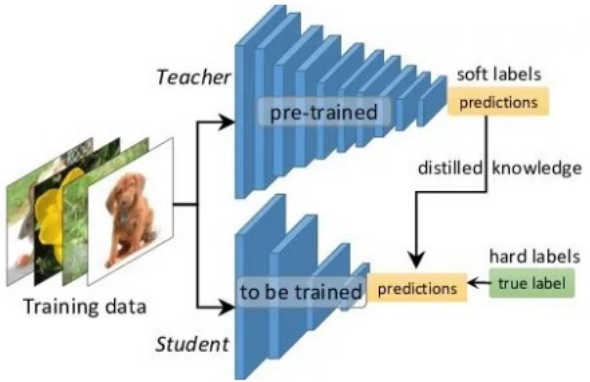
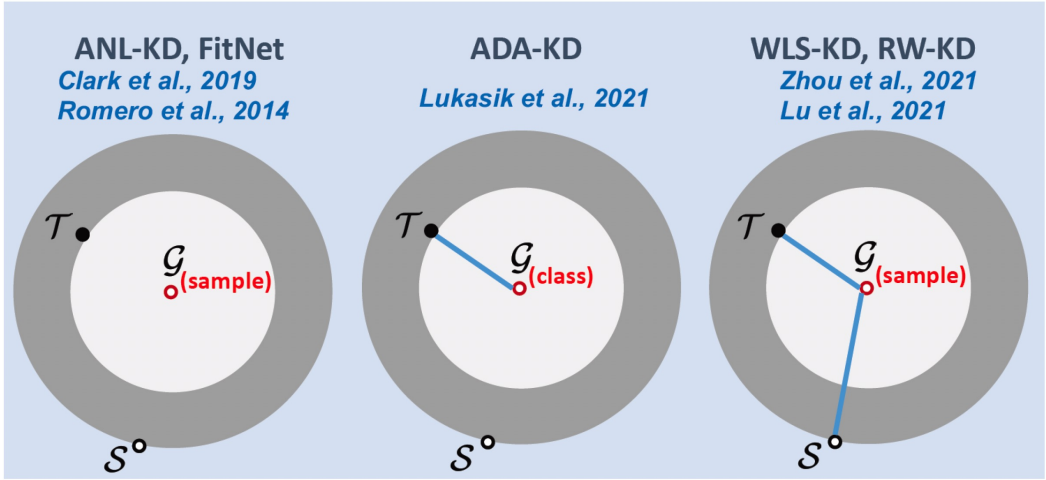


Image source: Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, Dark Knowledge.

Knowledge Fusion Ratio

- Current solutions [2-6] are sub-optimal:

Uniform fusion ratio across all samples / Cannot capture *full dynamics* of knowledge transfer.



Ignore the discrepancy between student's predictions *(S)* and teacher's predictions *(T)*, denoted as ***ST***.

- ***Research Question:***

How to design a better sample-wise ratio for knowledge trade-off?

Motivation Experiments

- ***Our claim:*** determining the knowledge fusion ratio depends on ***ST*** and the ***correctness of teacher's predictions.***
- Motivation experiments on CIFAR-100: a ResNet-34 teacher and a ResNet-18 student.

Step 1: partition the dataset into two subsets.

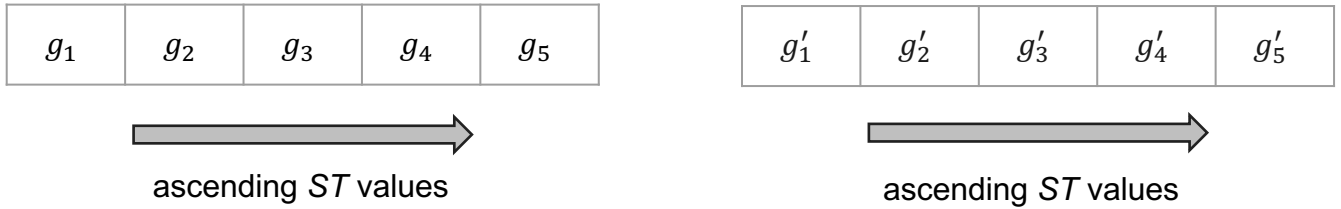
D: samples with correct teacher's predictions	*D':* samples with incorrect teacher's predictions

Step 2: receive preliminary knowledge through initial student training with $\alpha = 0.5$ over 50 epochs.

Step 3: compute ST across all samples (i.e., Euclidean distance between the student's and teacher's predicted class probabilities).

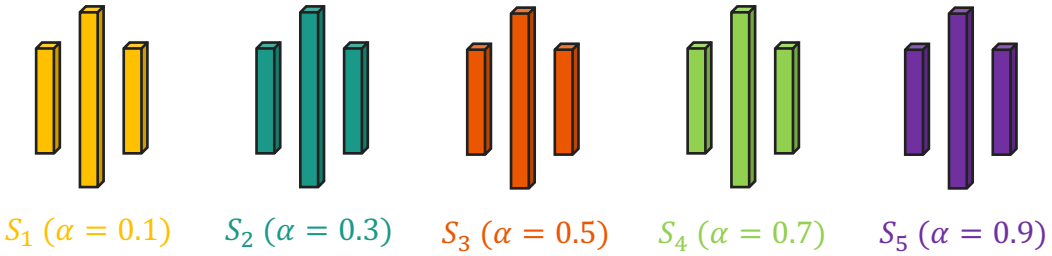
Motivation Experiments

Step 4: split D and D' into five equalized groups, respectively.



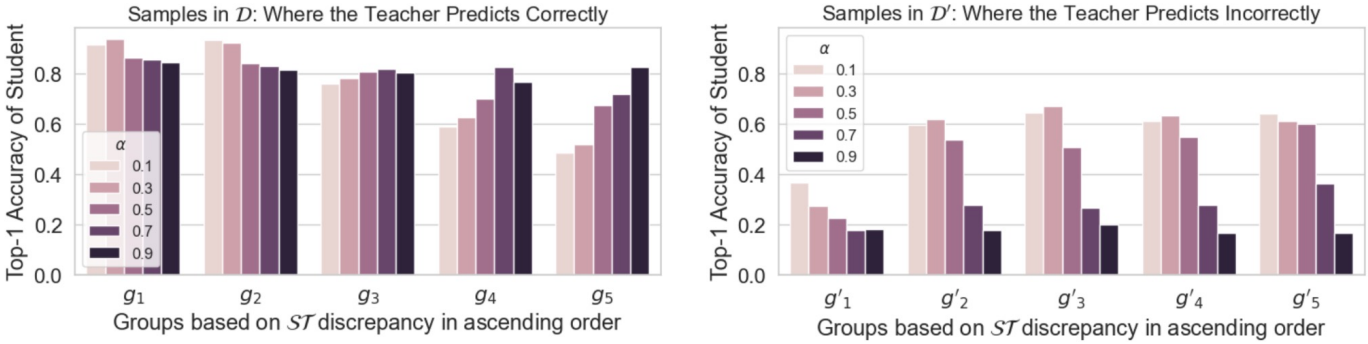
Step 5: further train the student with varying α values adjusted from {0.1, 0.3, 0.5, 0.7, 0.9}.

Five distinct students



Motivation Experiments

Step 6: evaluate these five students across all five g groups and five g' groups.



- ***Correct predictions:*** a higher *ST* indicates the higher learning potential from the teacher, favoring a larger α.
- ***Incorrect predictions:*** knowledge from the teacher is misleading, and thus a smaller α is advisable.
- Determining a proper sample-wise α relies on ***the teacher's or student's performances*** and ***the value of ST***.

Our TGeo-KD

TGeo-KD: learn the knowledge fusion ratio based on ***trilateral geometry*** within (S, T, G).

- Given a training sample (x_i, y_i), the knowledge fusion ratio is modeled as $\alpha_i = f_\omega(\Delta_i)$, where f_ω is one network parameterized by ω.
- Δ_i represents the unique geometric relations among (S_i, T_i, G_i).

Bilevel optimization: find the optimal sample-wise fusion ratio and the student network.

$$\min_{\omega} \mathcal{J}_{\text{val}}^{\text{outer}}(\theta^*(\omega)) = \frac{1}{N_{\text{val}}} \sum_{i=1}^{N_{\text{val}}} \mathcal{L}_i^{\text{GT}},$$

outer level: train the network f_ω parameterized by ω given fixed θ^*

$$\text{s.t. } \theta^*(\omega) = \underset{\theta}{\operatorname{argmin}}\, \mathcal{J}_{\text{train}}^{\text{inner}}(\theta, \omega) := \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} f_\omega(\Delta_i)\mathcal{L}_i^{\text{KD}} + \Big(1 - f_\omega(\Delta_i)\Big)\mathcal{L}_i^{\text{GT}}.$$

inner level: train the student parameterized by θ given fixed ω

Exploiting Trilateral Geometry

How to model the sample-wise trilateral geometry of Δ_i?

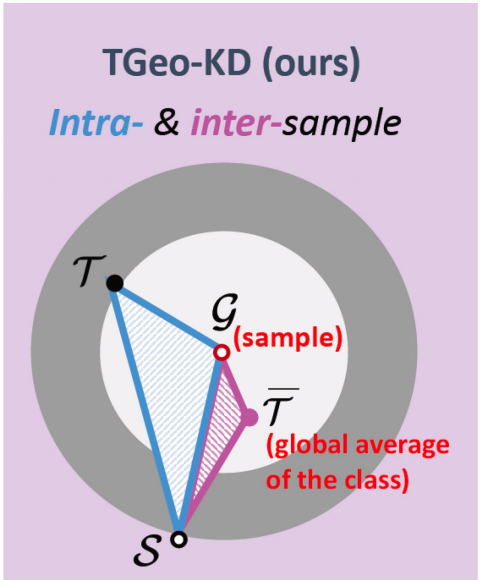
Intra-sample relations Δ_i^{STG}

- *three edges:* the student's correctness, the teacher's correctness, and the discrepancy between the student and teacher.

$$\mathbf{e}_i^{sg} := [\mathcal{G}_i - \mathcal{S}_i] \in \mathbb{R}^C, \mathbf{e}_i^{tg} := [\mathcal{G}_i - \mathcal{T}_i] \in \mathbb{R}^C, \mathbf{e}_i^{st} := [\mathcal{T}_i - \mathcal{S}_i] \in \mathbb{R}^C$$

- *three vertices:* the exact probability across all classes.

$$\Delta_i^{\mathcal{STG}} := [\mathbf{e}_i^{sg} \oplus \mathbf{e}_i^{tg} \oplus \mathbf{e}_i^{st} \oplus \mathcal{S}_i \oplus \mathcal{T}_i \oplus \mathcal{G}_i]$$

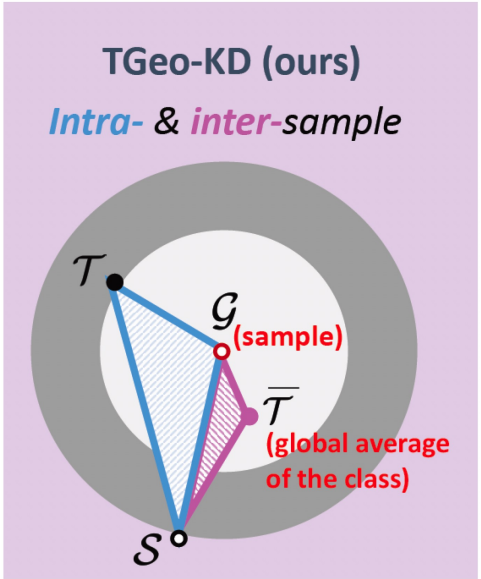


Exploiting Trilateral Geometry

Inter-sample relations $\Delta_i^{S\bar{T}G}$

- The teacher may perform poorly on outliers.
- Blindly using teacher predictions as the supervisory signal can result in the misleading knowledge.
- *An additional vertex* $\bar{T}_{c^i} \in \mathbb{R}^C$*:* the teacher's global average prediction on each class.
- *An additional triplet* $(S_i, \bar{T}_{c^i}, G_i)$*:* the trilateral geometry among inter-sample relations.

$$\Delta_i^{\mathcal{S}\bar{\mathcal{T}}\mathcal{G}} := [\mathbf{e}_i^{sg} \oplus \mathbf{e}_i^{\bar{t}g} \oplus \mathbf{e}_i^{s\bar{t}} \oplus \mathcal{S}_i \oplus \bar{\mathcal{T}}_{c^i} \oplus \mathcal{G}_i].$$



Experimental Results

Experiments on *three different tasks.*

- ***Image classification on CIFAR-100.***
- Consistent performance improvement when the architectural gap increases and hetero-architecture KD scenarios.

Table 1: Top-1 classification accuracy (%) on CIFAR-100. We re-implemented the methods denoted by * and calculated their average results (with standard deviation) over 5 repeated runs. For the remaining methods, we utilized the results provided or verified by the others (Tian et al., 2020; Zhou et al., 2021). The best performance is **bold**, while the second best is underlined.

Method	Same architecture style					Different architecture styles		
Teacher Student	WRN-40-2 WRN-40-1	ResNet-56 ResNet-20	ResNet-110 ResNet-32	ResNet-110 ResNet-20	ResNet-32×4 ResNet-8×4	ResNet-32×4 ShuffleNetV1	ResNet-32×4 ShuffleNetV2	WRN-40-2 ShuffleNetV1
Teacher	75.61	72.34	74.31	74.31	79.42	79.42	79.42	75.61
Student	71.98	69.06	71.14	69.06	72.50	70.50	71.82	70.50
FitNet	72.24	69.21	71.06	68.99	73.50	73.59	73.54	73.73
AT	72.77	70.55	72.31	70.22	73.44	71.73	72.73	73.32
SP	72.43	69.67	72.69	70.04	72.94	73.48	74.56	74.52
CC	72.21	69.63	71.48	69.48	72.97	71.14	71.29	71.38
VID	73.30	70.38	72.61	70.16	73.09	73.38	73.40	73.61
RKD	72.22	69.61	71.82	69.25	71.90	72.28	73.21	72.21
PKT	73.45	70.34	72.61	70.25	73.64	74.10	74.69	73.89
AB	72.38	69.47	70.98	69.53	73.17	73.55	74.31	73.34
FT	71.59	69.84	72.37	70.22	72.86	71.75	72.50	72.03
NST	72.24	69.60	71.96	69.53	73.30	74.12	74.68	74.89
CRD	74.14	71.16	73.48	71.46	75.51	75.11	75.65	76.05
Vanilla KD	73.54	70.66	73.08	70.67	73.33	74.07	74.45	74.83
ANL-KD*	72.81±0.25	72.13±0.18	72.50±0.21	72.28±0.21	75.07±0.26	72.58±0.23	73.11±0.14	75.27±0.32
ADA-KD*	74.67±0.19	72.22±0.21	73.19±0.12	72.29±0.27	75.78±0.34	71.45±0.16	72.20±0.24	75.49±0.28
WLS-KD	74.48	72.15	74.12	72.19	76.05	75.46	75.93	76.21
RW-KD*	73.92±0.22	70.33±0.26	71.78±0.15	71.24±0.16	74.86±0.29	70.45±0.25	70.69±0.17	74.15±0.29
TGeo-KD	**75.43**±0.16	**72.98**±0.14	**75.09**±0.13	**73.55**±0.20	**77.27**±0.25	**76.83**±0.17	**76.89**±0.14	**77.05**±0.23

Experimental Results

- ***Image classification on ImageNet.***
- Same architecture style: the improvement of *1.10%* over the strongest baseline.
- Hetero-architecture style: the improvement of *0.94%*.

Table 2: Top-1 and Top-5 classification accuracy on ImageNet. We re-implemented the methods denoted by * and used the author-provided or author-verified results for the others (Zhou et al., 2021).

Teacher: ResNet-34 → Student: ResNet-18		
Method	Top-1 ACC	Top-5 ACC
Teacher	73.31	91.42
Student	69.75	89.07
AT	71.03	90.04
NST	70.29	89.53
FSP	70.58	89.61
RKD	70.40	89.78
Overhaul	71.03	90.15
CRD	71.17	90.13
Vanilla KD	70.67	90.04
ANL-KD*	71.83±0.22	90.21±0.26
ADA-KD*	71.96±0.17	90.45±0.21
WLS-KD	72.04	90.70
RW-KD*	70.62±0.22	89.76±0.15
TGeo-KD	**72.89**±0.15	**91.80**±0.04

Teacher: ResNet-50 → Student: MobileNetV1		
Method	Top-1 ACC	Top-5 ACC
Teacher	76.16	92.87
Student	68.87	88.76
AT	70.18	89.68
FT	69.88	89.50
AB	68.89	88.71
RKD	68.50	88.32
Overhaul	71.33	90.33
CRD	69.07	88.94
Vanilla KD	70.49	89.92
ANL-KD*	70.40±0.15	89.25±0.22
ADA-KD*	71.08±0.24	90.17±0.16
WLS-KD	71.52	90.34
RW-KD*	70.15±0.16	89.40±0.19
TGeo-KD	**72.46**±0.14	**90.95**±0.17

Experimental Results

- ***Attack detection on HIL*** and ***click-through rate (CTR) prediction on Criteo.***

Table 3: Result comparison on HIL and Criteo under Teacher (12-layer BERT) → Student (4-layer BERT). The best performance is **bold**, while the second best is underlined. "⇑" indicates the metric value the higher the better, while "⇓" indicates the lower the better. Our TGeo-KD demonstrates a statistical significance for $p \leq 0.01$ compared to the strongest baseline based on the paired t-test.

Dataset	HIL			Criteo		
Method	ACC (%) ⇑	AUC (%) ⇑	NLL ⇓	ACC (%) ⇑	AUC (%) ⇑	NLL ⇓
Teacher	88.19	75.23	0.94	78.15	79.08	0.77
Student	87.64	67.58	1.02	69.43	69.02	1.79
Vanilla KD	87.55±0.56	69.52±0.70	1.00±0.04	71.08±0.48	69.42±0.60	1.51±0.05
ANL-KD	87.27±0.23	70.01±0.26	1.02±0.03	72.71±0.35	71.02±0.39	1.08±0.05
ADA-KD	90.15±0.34	70.02±0.21	0.99±0.02	72.15±0.33	71.01±0.35	1.15±0.04
WLS-KD	90.05±0.28	70.70±0.23	1.01±0.05	75.30±0.38	75.03±0.40	0.82±0.04
RW-KD	89.40±0.45	66.03±0.58	1.07±0.06	75.05±0.44	75.11±0.53	0.89±0.07
TGeo-KD	**92.39**±0.49	**71.65**±0.28	**0.94**±0.03	**77.80**±0.29	**77.00**±0.32	**0.81**±0.04

Conclusions

- An innovative method named ***TGeo-KD*** for learning sample-wise knowledge fusion ratios.
- Exploit ***the trilateral geometry*** among the supervision signals from the student, teacher, and ground truth by modeling both ***intra- and inter-sample geometric relations***.
- Comprehensive experiments to demonstrate the consistent superiority across ***diverse application domains***, as well as to highlight its adaptability across ***different architectures and model sizes***.

Reference

[1] Geoffrey Hinton, et al. “Distilling the knowledge in a neural network.” In *NeurIPS Deep Learning Workshop*, 2014.

[2] Kevin Clark, et al. “BAM! born-again multi-task networks for natural language understanding.” In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* pages 5931–5937, 2019.

[3] Adriana Romero, et al. “Fitnets: Hints for thin deep nets.” In *ICLR,* 2015.

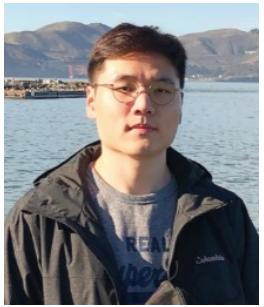
[4] Michal Lukasik, et al. “Teacher’s pet: understanding and mitigating biases in distillation.” In TMLR, 2022.

[5] Helong Zhou, et al. “Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective.” In *ICLR,* 2021.

[6] Peng Lu, et al. “RW-KD: Sample-wise loss terms re-weighting for knowledge distillation.” In *Findings of the Association for Computational Linguistics: EMNLP*, pages 3145–3152, 2021.

Thank you!

Please feel free to contact us if you have any questions:



Chengming Hu

Haolun Wu