



Full paper (ArXiv) Codes (Github)



# Logit Standardization in Knowledge Distillation

Shangquan Sun<sup>1,2</sup>, Wenqi Ren<sup>3</sup>, Jingzhi Li<sup>1</sup>, Rui Wang<sup>1,2</sup>, Xiaochun Cao<sup>3</sup>

<sup>1</sup>CAS, China <sup>2</sup>UCAS, China <sup>3</sup>Shenzhen Campus of SYSU, China



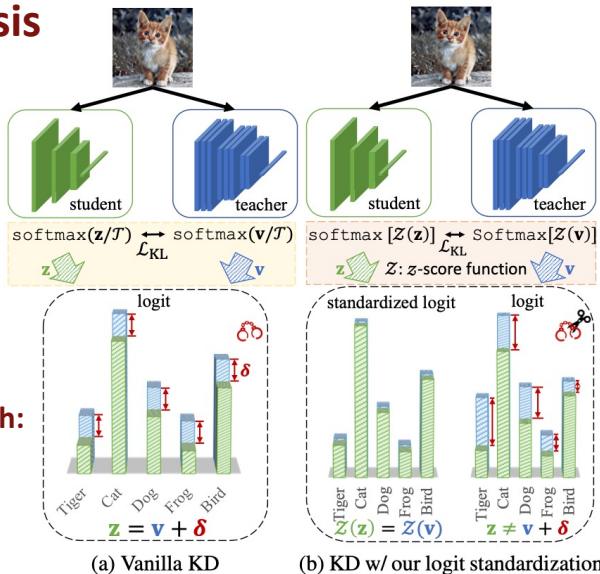
## Introduction & Analysis

- Common KD assumes  $\mathcal{T}_S = \mathcal{T}_T$  for all sample for simplicity
- But we find no explicit constraint on  $\mathcal{T}_S$  and  $\mathcal{T}_T$ , based on the derivation of softmax in KD by the entropy-maximum principle
- We find **2** issues when  $\mathcal{T}_S = \mathcal{T}_T$

### ISSUE 1

An implicit mandatory logit match:

- Given logit  $\mathbf{z}$  for  $S$  and  $\mathbf{v}$  for  $T$
- $\mathbf{z} = \mathbf{v} + \Delta$ , where  $\Delta = \bar{\mathbf{z}} - \bar{\mathbf{v}}$
- $\text{std}(\mathbf{z})/\text{std}(\mathbf{v}) = \mathcal{T}_S/\mathcal{T}_T = 1$

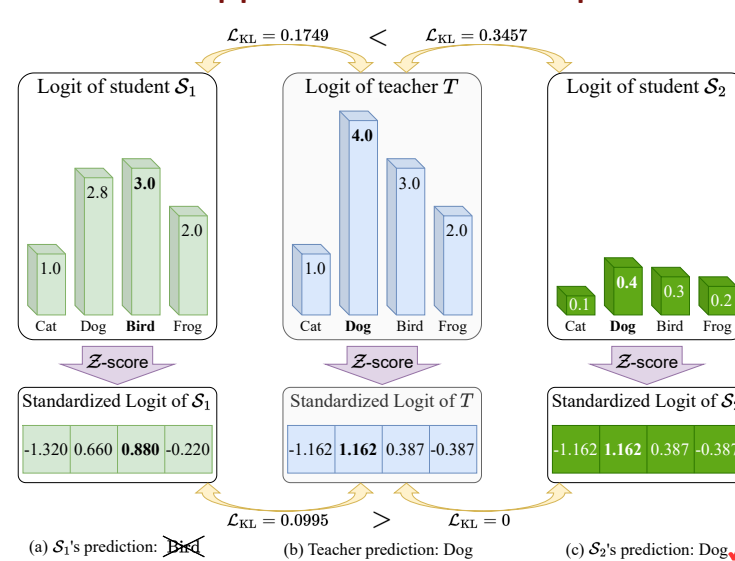


## Toy Case ISSUE 2

- Without ours:  $S_1$  has **better**  $\mathcal{L}_{KL}$  but **wrong** prediction
- $S_2$  has **worse**  $\mathcal{L}_{KL}$  but **correct** prediction

- With ours:  $S_2$  has **better**  $\mathcal{L}_{KL}$  and **correct** prediction
- Contradiction solved

Conventional KD pipeline fails to reflect student performance



## Proposed Method: Logit Standardization

- Determine Temperature adaptively based on weighted  $\mathcal{Z}$ -score
- Serve as a beneficial pre-process for the existing logit-based KD

**Algorithm 1: Weighted  $\mathcal{Z}$ -score function.**

**Input:** Input vector  $\mathbf{x}$  and Base temperature  $\tau$

**Output:** Standardized vector  $\mathcal{Z}(\mathbf{x}; \tau)$

- 1  $\bar{\mathbf{x}} \leftarrow \frac{1}{K} \sum_{k=1}^K \mathbf{x}^{(k)}$
- 2  $\sigma(\mathbf{x}) \leftarrow \sqrt{\frac{1}{K} \sum_{k=1}^K (\mathbf{x}^{(k)} - \bar{\mathbf{x}})^2}$
- 3 **return**  $(\mathbf{x} - \bar{\mathbf{x}})/\sigma(\mathbf{x})/\tau$

**Algorithm 2:  $\mathcal{Z}$ -score logit standardization pre-process in knowledge distillation.**

**Input:** Transfer set  $\mathcal{D}$  with image-label sample pair  $\{\mathbf{x}_n, y_n\}_{n=1}^N$ , Base Temperature  $\tau$ , Teacher  $f_T$ , Student  $f_S$ , Loss  $\mathcal{L}_{KD}$  (e.g.,  $\mathcal{L}_{KL}$ ), loss weight  $\lambda$ , and  $\mathcal{Z}$ -score function  $\mathcal{Z}$  in Algo. 1

**Output:** Trained student model  $f_S$

- 1 **foreach**  $(\mathbf{x}_n, y_n)$  in  $\mathcal{D}$  **do**
- 2  $\mathbf{v}_n \leftarrow f_T(\mathbf{x}_n), \mathbf{z}_n \leftarrow f_S(\mathbf{x}_n)$
- 3  $q(\mathbf{v}_n) \leftarrow \text{softmax}[\mathcal{Z}(\mathbf{v}_n; \tau)]$
- 4  $q(\mathbf{z}_n) \leftarrow \text{softmax}[\mathcal{Z}(\mathbf{z}_n; \tau)]$
- 5  $q'(\mathbf{z}_n) \leftarrow \text{softmax}(\mathbf{z}_n)$
- 6 Update  $f_S$  towards minimizing  $\lambda_{CE} \mathcal{L}_{CE}(y_n, q'(\mathbf{z}_n)) + \lambda_{KD} \tau^2 \mathcal{L}(q(\mathbf{v}_n), q(\mathbf{z}_n))$
- 7 **end**

- Four Beneficial properties of standardized logit:

1. Zero mean
2. Finite logit std.
3. Monotonicity
4. Boundedness within  $\left[-\frac{\sqrt{K-1}}{\tau}, \frac{\sqrt{K-1}}{\tau}\right]$

## Experiments

### Distillation on CIFAR-100

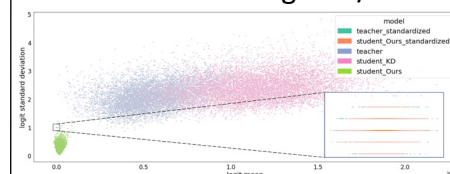
Part of Table for Different Structures				Part of Table for Identical Structures				
Type	Teacher	ResNet32×4	ResNet32×4	ResNet32×4	WRN-40-2	ResNet56	ResNet110	ResNet110
	Student	SHN-V2	WRN-16-2	WRN-40-2	WRN-16-2	ResNet20	ResNet32	ResNet20
Feature	FitNet [31]	73.54	74.70	77.69	73.58	69.21	71.06	68.99
	AT [46]	72.73	73.91	77.43	74.08	70.55	72.31	70.65
	RKD [29]	73.21	74.86	77.82	73.35	69.61	71.82	69.25
	CRD [37]	75.65	75.65	78.15	75.48	71.16	73.48	71.46
	OFD [12]	76.82	76.17	79.25	75.24	70.98	73.23	71.29
	ReviewKD [5]	77.78	76.11	78.96	76.12	71.89	73.89	71.34
Logit	SimKD [4]	78.39	77.17	79.29	75.53	71.05	73.92	71.06
	CAT-KD [10]	78.41	76.97	78.59	75.60	71.62	73.62	71.37
	KD [13]	74.45	74.90	77.70	74.92	70.66	73.08	70.67
	KD+Ours	75.56	75.26	77.92	76.11	71.43	74.17	71.48
	$\Delta$	1.11	0.36	0.22	1.19	0.77	1.09	0.81
	CTKD [24]	75.37	74.57	77.66	75.45	71.19	73.52	70.99
Logit	CTKD+Ours	76.18	75.16	77.99	76.08	71.34	74.01	71.39
	$\Delta$	0.81	0.59	0.33	0.63	0.15	0.49	0.40
	DKD [50]	77.07	75.70	78.46	76.24	71.97	74.11	71.06
	DKD+Ours	77.37	76.19	78.95	76.39	72.32	74.29	71.85
	$\Delta$	0.30	0.49	0.49	0.15	0.35	0.18	0.79
	MLKD [17]	78.44	76.52	79.26	76.63	72.19	74.11	71.89
Logit	MLKD+Ours	78.76	77.53	79.66	76.95	72.33	74.32	72.27
	$\Delta$	0.32	1.01	0.40	0.32	0.14	0.21	0.38

### Distillation on ImageNet

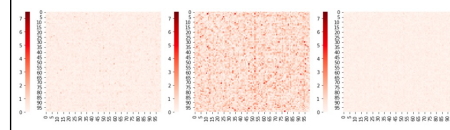
Teacher/Student	ResNet34/ResNet18	ResNet50/MN-V1
Accuracy	top-1 top-5	top-1 top-5
Teacher	73.31 91.42	76.16 92.86
Student	69.75 89.07	68.87 88.76
AT [46]	70.69 90.01	69.56 89.33
OFD [12]	70.81 89.98	71.25 90.34
CRD [37]	71.17 90.13	71.37 90.41
ReviewKD [5]	71.61 90.51	72.56 91.00
SimKD [4]	71.59 90.48	72.25 90.86
CAT-KD [10]	71.26 90.45	72.24 91.13
KD [13]	71.03 90.05	70.50 89.80
KD+Ours	71.42+0.39 90.29+0.24	72.18+1.68 90.80+1.00
KD+CTKD [24]	71.38 90.27	71.16 90.11
KD+CTKD+Ours	71.81+0.43 90.46+0.19	72.92+1.76 91.25+1.14
DKD [50]	71.70 90.41	72.05 91.05
DKD+Ours	71.88+0.18 90.58+0.17	72.85+0.80 91.23+0.18
MLKD [17]	71.90 90.55	73.01 91.42
MLKD+Ours	72.08+0.18 90.74+0.19	73.22+0.21 91.59+0.17

### Visualization

- No restriction on mean and std.
- Better match of logits w/ ours



Bivariate histogram of logit mean and std.



(a) Vanilla KD (b) Ours w/o  $\mathcal{Z}$ -score (c) Ours w/  $\mathcal{Z}$ -score  
Mean: 0.27, Max: 3.03. Mean: 0.94, Max: 7.36. Mean: 0.18, Max: 1.18.  
Heatmap of avg. logit diff. between  $T$  &  $S$