Logit Standardization in Knowledge Distillation

Shangquan Sun[1,2], Wenqi Ren[3]†, Jingzhi Li[1], Rui Wang[1,2], Xiaochun Cao[3]

CVPR 2024
(Highlight)

- Problem/Objective
  - Knowledge Distillation
  - Classification Task

- Contribution/Key Idea
  - Prove the irrelevance between T-S temperature
  - Show the drawbacks of the shared Temperature
  - Propose Logit Standardization

Shangquan Sun

University of Chinese Academy of Sciences
在 iie.ac.cn 的电子邮件经过验证 - 首页
Computer Vision Machine Learning

标题

Logit Standardization in Knowledge Distillation
S Sun, W Ren, J Li, R Wang, X Cao
CVPR 2024 (Highlight)

Rethinking image restoration for object detection
S Sun, W Ren, T Wang, X Cao
NeurIPS 2022

Event-aware video deraining via multi-patch progressive learning
S Sun, W Ren, J Li, K Zhang, M Liang, X Cao
IEEE Transactions on Image Processing

Restoring Images in Adverse Weather Conditions via Histogram Transformer
S Sun, W Ren, X Gao, R Wang, X Cao
ECCV 2024

DI-Retinex: Digital-Imaging Retinex Theory for Low-Light Image Enhancement
S Sun, W Ren, J Peng, F Song, X Cao
arXiv preprint arXiv:2404.03327

EnsIR: An Ensemble Algorithm for Image Restoration via Gaussian Mixture Models
S Sun, W Ren, Z Liu, H Park, R Wang, X Cao
NeurIPS 2024

김범준

CVPR 2024
(Highlight)

Logit Standardization in Knowledge Distillation

Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

- **KD에 대해**

$$\mathcal{L}_{KD} = \alpha \times \mathcal{L}_{KL}(y, q(z)) + \beta \times \mathcal{L}_{KL}(q(v;\tau), q(z;\tau))$$

GT → Student: $\mathcal{L}_{KL}(y, q(z))$

Teacher → Student: $\mathcal{L}_{KL}(q(v;\tau), q(z;\tau))$

- Distillation through **Softmax q + Temperature T to soften pseudo-label**
- Temperature은 softmax 시 확률 분포를 조정하는 값(>1이면 고르게, <1이면 차이 극대화)

[1] Hinton, Geoffrey. "Distilling the Knowledge in a Neural Network." *NeurIPS 2014*

[Logit Standardization in Knowledge Distillation]

Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

CVPR 2024
(Highlight)

- **KD에 대해**

$$\mathcal{L}_{KD} = \alpha \times \mathcal{L}_{KL}(y, q(z)) + \beta \times \mathcal{L}_{KL}(q(v;\tau), q(z;\tau))$$

GT → Student: $\mathcal{L}_{KL}(y, q(z))$

Teacher → Student: $\mathcal{L}_{KL}(q(v;\tau), q(z;\tau))$

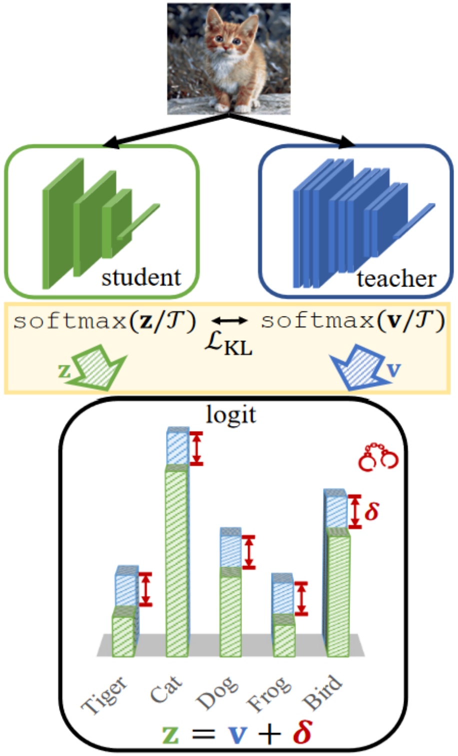
- Temperature T가 공유되는 문제점 (Teacher & Student)

김범준

[Logit Standardization in Knowledge Distillation](#)
Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

CVPR 2024
(Highlight)

- **KD에 대해**



(a) Vanilla KD

- Student는 독자적인 capability 존재
- Teacher = Student + δ 로 차이가 일정한 관계 (* δ 가 아님!!!)

  → 만약 * δ 였다면 같은 scale의 Temperature로 나눠주는 것이 Ok

- 그래서 Logit 에 Bias가 있는거는 현재 Hinton KD 수식으로 설명 불가

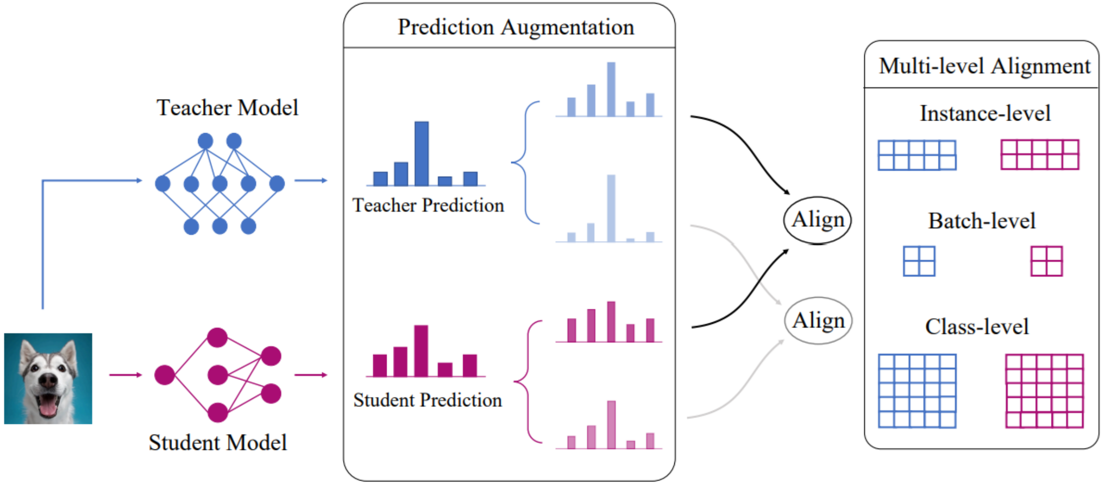
김범준

CVPR 2024 (Highlight)

[Logit Standardization in Knowledge Distillation](#)

Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

- **Related Work**



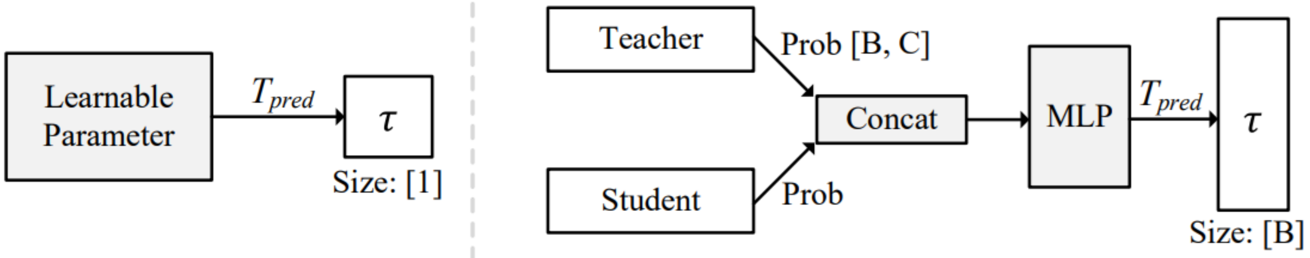
- 여러 Temperature (=Level) 에서 distillation 해보면 더 좋다.

[1]Jin, Ying, Jiaqi Wang, and Dahua Lin. "Multi-level logit distillation, *CVPR 2023.*

김범준

Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

- **Related Work**



(a) Global Temperature (b) Instance-wise Temperature

- Sample에 따라서 Temperature를 다르게 학습해서, 각각의 sample에 맞는 Temperature을 주자.
  → 여전히 Teacher / Student가 Temperature 공유하는 문제점

[1]Li, Zheng, et al. "Curriculum temperature for knowledge distillation." *AAAI 2023*

[Logit Standardization in Knowledge Distillation]

Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

CVPR 2024 (Highlight)

- **Related Work**

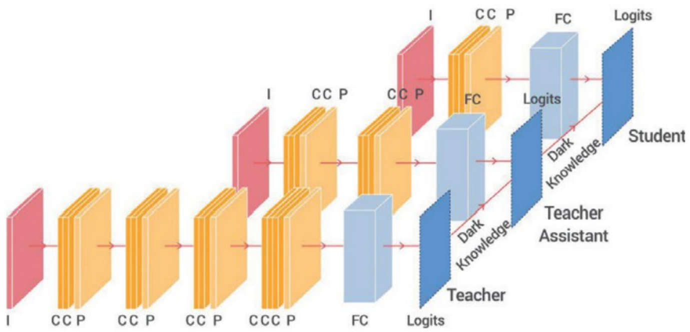


Figure 1: TA fills the gap between student & teacher

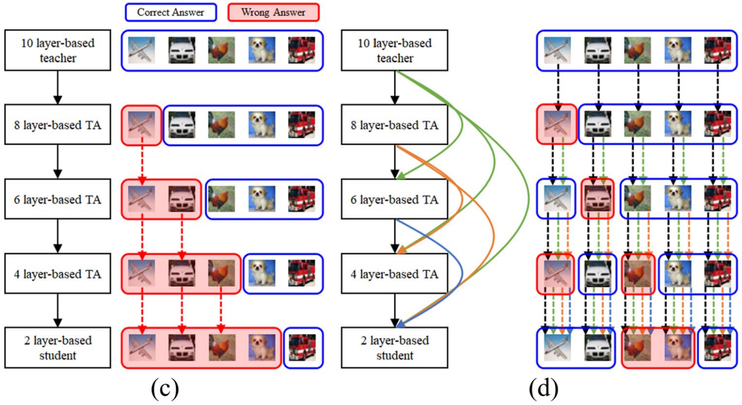


Figure 1. **Problem definition of the large gap between a teacher and a student network.** (a) In general, the difference between

- Teacher - TA - Student 를 이용해서 Gap을 줄이려는 노력

  → But, 수식적으로 완벽하진 x

[1] Mirzadeh, Seyed Iman, et al. "Improved knowledge distillation via teacher assistant." *AAAI 2020*
[2] Son, Wonchul, et al. "Densely guided knowledge distillation using multiple teacher assistants." *CVPR 2021*

[Logit Standardization in Knowledge Distillation](#)

Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

CVPR 2024
(Highlight)

- ## Background and Notation

K-class classification task with N samples $\{\mathbf{x}_n, y_n\}_{n=1}^{N}$.

Each $\mathbf{x}_n \in \mathbb{R}^{H\times W}$ is an input and y_n is its corresponding output.

Teacher f_T and student f_S produce logits $\mathbf{z}_n = f_S(\mathbf{x}_n)$ and $\mathbf{v}_n = f_T(\mathbf{x}_n)$, respectively, for n^{th} sample.

Softmax function $q(\mathbf{z}_n)^{(k)} = \frac{\exp\left(\mathbf{z}_n^{(k)}/\tau\right)}{\sum_{m=1}^{k}\exp\left(\mathbf{z}_n^{(m)}/\tau\right)}, q(\mathbf{v}_n)^{(k)} = \frac{\exp\left(\mathbf{v}_n^{(k)}/\tau\right)}{\sum_{m=1}^{k}\exp\left(\mathbf{v}_n^{(m)}/\tau\right)}$ with temperature τ is general.

Knowledge distillation essentially aims to let $q(\mathbf{z}_n)^{(k)}$ to mimic $q(\mathbf{v}_n)^{(k)}$.

$$\mathcal{L}_{\mathrm{KL}}\left(q(\mathbf{z}_n)\|q(\mathbf{v}_n)\right) = \sum_{k=1}^{K} q(\mathbf{z}_n)^{(k)} \log\left(\frac{q(\mathbf{v}_n)^{(k)}}{q(\mathbf{z}_n)^{(k)}}\right).$$

$\mathcal{L}_{\mathrm{KL}}$ is “theoretically” equivalent to $\mathcal{L}_{\mathrm{CE}}\left(q(\mathbf{z}_n), q(\mathbf{v}_n)\right) = -\sum_{k=1}^{K} q(\mathbf{z}_n)^{(k)} \log\left(q(\mathbf{z}_n)^{(k)}\right)$.

[Logit Standardization in Knowledge Distillation]
Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

- **Methodology - Derivation of Softmax (Classification)**

$$\max_{q} \mathcal{L}_1 = -\sum_{n=1}^{N}\sum_{k=1}^{K} q(\mathbf{v}_n)^{(k)} \log q(\mathbf{v}_n)^{(k)}$$ → Classification의 objective function은 entropy를 maximize 하는 문제로 볼 수 있음

- **제약 조건**

$$s.t. \begin{cases} \sum_{k=1}^{K} q(\mathbf{v}_n)^{(k)} = 1, \quad \forall n \\ \mathbb{E}_q[\mathbf{v}_n] = \sum_{k=1}^{K} \mathbf{v}_n^{(k)} q(\mathbf{v}_n)^{(k)} = \mathbf{v}_n^{(y_n)}, \quad \forall n. \end{cases}$$

→ q(Vn)이 probability distribution이 되게끔

→ q(Vn)을 target one-hot vector (g^n)이 되게끔

the target class. Suppose $\hat{q}_n$ to be the one-hot hard probability distribution whose values are all zero except at the target index $\hat{q}_n^{(y_n)} = 1$. The second constraint is then actu-

[Logit Standardization in Knowledge Distillation]

Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

- **Methodology - Derivation of Softmax (Classification)**

$$\mathcal{L}_T = \mathcal{L}_1 + \sum_{n=1}^{N} \alpha_{1,n} \left(\sum_{k=1}^{K} q(\mathbf{v}_n)^{(k)} - 1 \right) + \sum_{n=1}^{N} \alpha_{2,n} \left(\sum_{k=1}^{K} \mathbf{v}_n^{(k)} q(\mathbf{v}_n)^{(k)} - \mathbf{v}_n^{(y_n)} \right)$$

→ 라그랑지안 multipliers로 미분해서 최대화하는 방향 찾기

Softmax function $q(\mathbf{z}_n)^{(k)} = \frac{\exp(\mathbf{z}_n^{(k)}/\tau)}{\sum_{m=1}^{k} \exp(\mathbf{z}_n^{(m)}/\tau)}$

↓ 편미분

$$\frac{\partial \mathcal{L}_T}{\partial q(\mathbf{v}_n)^{(k)}} = -1 - \log q(\mathbf{v}_n)^{(k)} + \alpha_{1,n} + \alpha_{2,n} \mathbf{v}_n^{(k)}$$

$$q(\mathbf{v}_n)^{(k)} = \exp\left(\alpha_{2,n} \mathbf{v}_n^{(k)}\right) / Z_T$$

→ softmax를 쓰는게 최적의 q(Vn)

$$Z_T = \exp(1 - \alpha_{1,n}) = \sum_{m=1}^{K} \exp\left(\alpha_{2,n} \mathbf{v}_n^{(m)}\right)$$

[Logit Standardization in Knowledge Distillation](#)

Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

CVPR 2024
(Highlight)

- ## **Methodology - Derivation of Softmax (KD)**

$$\max_q \mathcal{L}_2 = -\sum_{n=1}^{N}\sum_{k=1}^{K} q(\mathbf{z}_n)^{(k)} \log q(\mathbf{z}_n)^{(k)}$$

$$\text{s.t.} \begin{cases} \sum_{k=1}^{K} q(\mathbf{z}_n)^{(k)} = 1, & \forall n \\ \sum_{k=1}^{K} \mathbf{z}_n^{(k)} q(\mathbf{z}_n)^{(k)} = \mathbf{z}_n^{(y_n)}, & \forall n \\ \sum_{k=1}^{K} \mathbf{z}_n^{(k)} q(\mathbf{z}_n)^{(k)} = \sum_{k=1}^{K} \mathbf{z}_n^{(k)} q(\mathbf{v}_n)^{(k)}, & \forall n. \end{cases}$$

$$\begin{aligned} \mathcal{L}_S = \mathcal{L}_2 &+ \sum_{n=1}^{N} \beta_{1,n}\left(\sum_{k=1}^{K} q(\mathbf{z}_n)^{(k)} - 1\right) \\ &+ \sum_{n=1}^{N} \beta_{2,n}\left(\sum_{k=1}^{K} \mathbf{z}_n^{(k)} q(\mathbf{z}_n)^{(k)} - \mathbf{z}_n^{(y_n)}\right) \\ &+ \sum_{n=1}^{N} \beta_{3,n} \sum_{k=1}^{K} \mathbf{z}_n^{(k)} \left(q(\mathbf{z}_n)^{(k)} - q(\mathbf{v}_n)^{(k)}\right) \end{aligned}$$

$$\frac{\partial \mathcal{L}_S}{\partial q(\mathbf{z}_n)^{(k)}} = -1 - \log q(\mathbf{z}_n)^{(k)} + \beta_{1,n} + \beta_{2,n}\mathbf{z}_n^{(k)} + \beta_{3,n}\mathbf{z}_n^{(k)}$$

$$q(\mathbf{z}_n)^{(k)} = \exp\left(\beta_n \mathbf{z}_n^{(k)}\right) / Z_S \rightarrow$$ softmax를 쓰는게 최적의 q(Zn)

$$Z_S = \exp(1 - \beta_{1,n}) = \sum_{k=1}^{K} \exp(\beta_n \mathbf{z}_n^{(k)})$$

김범준

Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

- # Methodology - Derivation of Softmax

$$\underset{q(\mathbf{v}_n)^{(k)}}{\operatorname{argmax}} \mathcal{L}_T = \frac{\exp\left(\alpha_{2,n}\mathbf{v}_n^{(k)}\right)}{\sum_{m=1}^{K}\exp\left(\alpha_{2,n}\mathbf{v}_n^{(m)}\right)} = \operatorname{softmax}\left(\mathbf{v}_n^{(k)}; \alpha_{2,n}^{-1}\right)$$

$$\underset{q(\mathbf{z}_n)^{(k)}}{\operatorname{argmax}} \mathcal{L}_S = \frac{\exp\left(\beta_n\mathbf{z}_n^{(k)}\right)}{\sum_{m=1}^{K}\exp\left(\beta_n\mathbf{z}_n^{(m)}\right)} = \operatorname{softmax}\left(\mathbf{z}_n^{(k)}; \beta_n^{-1}\right)$$

Optimal q equivalents to softmax.

$\beta_n = \alpha_{2,n} = 1/\mathcal{T}$ → 일반적인 KD에서 쓰는 Temperature 식

$\beta_n = \alpha_{2,n} = 1$ → 일반적인 Classification에서 쓰는 softmax 식

+ 알파, 베타 달라도 된다 !! Temperature 달라도 된다!!

Logit Standardization in Knowledge Distillation

Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

CVPR 2024 (Highlight)

- **Drawback of shared Temperature**

$$q\left(\mathbf{z}_n; a_S, b_S\right)^{(k)} = \frac{\exp\left[\left(\mathbf{z}_n^{(k)} - a_S\right)/b_S\right]}{\sum_{m=1}^{K} \exp\left[\left(\mathbf{z}_n^{(m)} - a_S\right)/b_S\right]}$$

→ Softmax를 다음과 같은 식으로 가정 (a,b의 bias가 추가된)

- Knowledge distillation에선 Zn, Vn에 softmax했을때 같이 지게끔 하면된다

$$\frac{\exp\left[\left(\mathbf{z}_n^{(i)} - a_S\right)/b_S\right]}{\exp\left[\left(\mathbf{z}_n^{(j)} - a_S\right)/b_S\right]} = \frac{\exp\left[\left(\mathbf{v}_n^{(i)} - a_T\right)/b_T\right]}{\exp\left[\left(\mathbf{v}_n^{(j)} - a_T\right)/b_T\right]} \Rightarrow \left(\mathbf{z}_n^{(i)} - \mathbf{z}_n^{(j)}\right)/b_S = \left(\mathbf{v}_n^{(i)} - \mathbf{v}_n^{(j)}\right)/b_T$$

$$\left(\mathbf{z}_n^{(i)} - \overline{\mathbf{z}}_n\right)/b_S = \left(\mathbf{v}_n^{(i)} - \overline{\mathbf{v}}_n\right)/b_T \rightarrow \frac{\sigma(\mathbf{z}_n)^2}{\sigma(\mathbf{v}_n)^2} = \frac{\frac{1}{K}\sum_{i=1}^{K}\left(\mathbf{z}_n^{(i)} - \overline{\mathbf{z}}_n\right)^2}{\frac{1}{K}\sum_{i=1}^{K}\left(\mathbf{v}_n^{(i)} - \overline{\mathbf{v}}_n\right)^2} = \frac{b_S^2}{b_T^2}$$

1 to K 합

김범준

Logit Standardization in Knowledge Distillation

Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

CVPR 2024
(Highlight)

- **Drawback of shared Temperature**

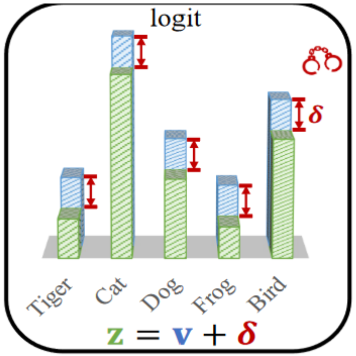
$$\frac{\sigma(\mathbf{z}_n)^2}{\sigma(\mathbf{v}_n)^2} = \frac{\frac{1}{K}\sum_{i=1}^{K}\left(\mathbf{z}_n^{(i)} - \overline{\mathbf{z}}_n\right)^2}{\frac{1}{K}\sum_{i=1}^{K}\left(\mathbf{v}_n^{(i)} - \overline{\mathbf{v}}_n\right)^2} = \frac{b_S^2}{b_T^2}$$

$$(b_S = b_T)$$

즉 if) Temperature 같다면 2가지 문제점

1) $\mathbf{z}_n^{(i)} = \mathbf{v}_n^{(i)} + \Delta_n,$

2) $\sigma(\mathbf{z}_n) = \sigma(\mathbf{v}_n)$



This is another shackle applied to student restricting the standard deviation of its predicted logits. In contrast, since

결론) Teacher - Student 서로 temperature 다르게 해야한다.

김범준

[Logit Standardization in Knowledge Distillation](#)

Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

CVPR 2024
(Highlight)

- **propose method**

  - Z-score 정규화

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

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

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

1 $\overline{\mathbf{x}} \leftarrow \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}^{(k)}$

2 $\sigma(\mathbf{x}) \leftarrow \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\mathbf{x}^{(k)}-\overline{\mathbf{x}}\right)^2}$

3 **return** $(\mathbf{x}-\overline{\mathbf{x}})/\sigma(\mathbf{x})/\tau$

정규분포 Z score $$Z = \frac{x-\mu}{\sigma}$$

본 논문 Z-score $$Z(x;\tau) = \frac{x-\bar{x}}{\sigma(x)\cdot\tau}$$

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 τ, Teacher f_T, Student f_S, Loss $\mathcal{L}_{\mathrm{KD}}$ (e.g., $\mathcal{L}_{\mathrm{KL}}$), loss weight λ, 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 $\quad \mathbf{v}_n \leftarrow f_T(\mathbf{x}_n)$, $\mathbf{z}_n \leftarrow f_S(\mathbf{x}_n)$

3 $\quad q(\mathbf{v}_n) \leftarrow \mathtt{softmax}\left[\mathcal{Z}(\mathbf{v}_n;\tau)\right]$

4 $\quad q(\mathbf{z}_n) \leftarrow \mathtt{softmax}\left[\mathcal{Z}(\mathbf{z}_n;\tau)\right]$

5 $\quad q'(\mathbf{z}_n) \leftarrow \mathtt{softmax}\left(\mathbf{z}_n\right)$

6 $\quad$ Update f_S towards minimizing $\lambda_{\mathrm{CE}}\mathcal{L}_{\mathrm{CE}}\left(y_n, q'(\mathbf{z}_n)\right) + \lambda_{\mathrm{KD}}\tau^2\mathcal{L}\left(q(\mathbf{v}_n), q(\mathbf{z}_n)\right)$

7 **end**

김범준

[Logit Standardization in Knowledge Distillation]()

Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

CVPR 2024
(Highlight)

- **Toy case**

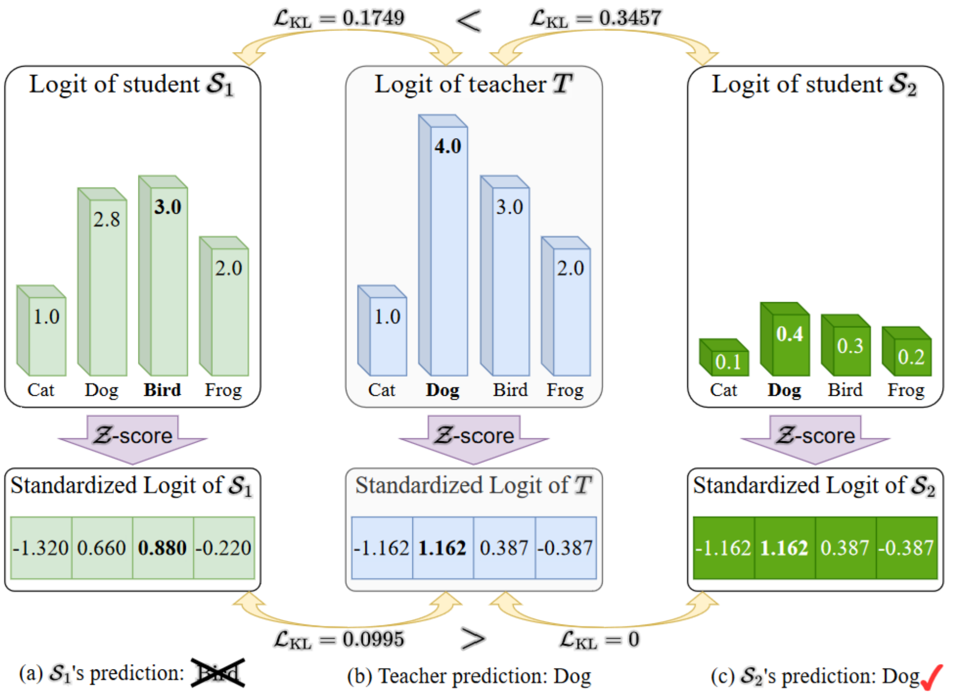


Figure 2. A toy case where two students, $\mathcal{S}_1$ and $\mathcal{S}_2$, learning

[Logit Standardization in Knowledge Distillation](#)

Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

CVPR 2024
(Highlight)

- # Experiment - CIFAR 100

Table 1. The Top-1 Accuracy (%) of different knowledge distillation methods on the validation set of CIFAR-100 [18]. The teacher and student have distinct architectures. The KD methods are sorted by the types, i.e., feature-based and logit-based. We apply our logit standardization to the existing logit-based methods and use Δ to show its performance gain. The values in blue denote slight enhancement and those in red non-trivial enhancement no less than 0.15. The best and second best results are emphasized in **bold** and underlined cases.

Type	Teacher	ResNet32×4	ResNet32×4	ResNet32×4	WRN-40-2	WRN-40-2	VGG13	ResNet50
		79.42	79.42	79.42	75.61	75.61	74.64	79.34
	Student	SHN-V2	WRN-16-2	WRN-40-2	ResNet8×4	MN-V2	MN-V2	MN-V2
		71.82	73.26	75.61	72.50	64.60	64.60	64.60
Feature	FitNet [33]	73.54	74.70	77.69	74.61	68.64	64.16	63.16
	AT [53]	72.73	73.91	77.43	74.11	60.78	59.40	58.58
	RKD [31]	73.21	74.86	77.82	75.26	69.27	64.52	64.43
	CRD [39]	75.65	75.65	78.15	75.24	70.28	69.73	69.11
	OFD [12]	76.82	76.17	79.25	74.36	69.92	69.48	69.04
	ReviewKD [5]	77.78	76.11	78.96	74.34	<u>71.28</u>	70.37	69.89
	SimKD [4]	78.39	<u>77.17</u>	<u>79.29</u>	75.29	70.10	69.44	69.97
	CAT-KD [10]	78.41	76.97	78.59	75.38	70.24	69.13	**71.36**
Logit	KD [13]	74.45	74.90	77.70	73.97	68.36	67.37	67.35
	KD+Ours	75.56	75.26	77.92	77.11	69.23	68.61	69.02
	Δ	1.11	0.36	0.22	3.14	0.87	1.24	1.67
	CTKD [24]	75.37	74.57	77.66	74.61	68.34	68.50	68.67
	CTKD+Ours	76.18	75.16	77.99	77.03	69.53	68.98	69.36
	Δ	0.81	0.59	0.33	2.42	1.19	0.48	0.69
	DKD [57]	77.07	75.70	78.46	75.56	69.28	69.71	70.35
	DKD+Ours	77.37	76.19	78.95	76.75	70.01	69.98	70.45
	Δ	0.30	0.49	0.49	1.19	0.73	0.27	0.10
	MLKD [17]	<u>78.44</u>	76.52	79.26	<u>77.33</u>	70.78	<u>70.57</u>	71.04
	MLKD+Ours	**78.76**	**77.53**	**79.66**	**77.68**	**71.61**	**70.94**	<u>71.19</u>
	Δ	0.32	1.01	0.40	0.35	0.83	0.37	0.15

Table 2. The Top-1 Accuracy (%) of different knowledge distillation methods on the validation set of CIFAR-100 [18]. The teacher and student have identical architectures but different configurations. The KD methods are sorted by the types. We apply our logit standardization to the existing logit-based methods and use Δ to show its performance gain. The values in blue denote slight enhancement and those in red non-trivial enhancement no less than 0.15. The best and second best results are emphasized in **bold** and underlined cases.

Type	Teacher	ResNet32×4	VGG13	WRN-40-2	WRN-40-2	ResNet56	ResNet110	ResNet110
		79.42	74.64	75.61	75.61	72.34	74.31	74.31
	Student	ResNet8×4	VGG8	WRN-40-1	WRN-16-2	ResNet20	ResNet32	ResNet20
		72.50	70.36	71.98	73.26	69.06	71.14	69.06
Feature	FitNet [33]	73.50	71.02	72.24	73.58	69.21	71.06	68.99
	AT [53]	73.44	71.43	72.77	74.08	70.55	72.31	70.65
	RKD [31]	71.90	71.48	72.22	73.35	69.61	71.82	69.25
	CRD [39]	75.51	73.94	74.14	75.48	71.16	73.48	71.46
	OFD [12]	74.95	73.95	74.33	75.24	70.98	73.23	71.29
	ReviewKD [5]	75.63	74.84	75.09	76.12	71.89	73.89	71.34
	SimKD [4]	<u>78.08</u>	74.89	74.53	75.53	71.05	73.92	71.06
	CAT-KD [10]	76.91	74.65	74.82	75.60	71.62	73.62	71.37
Logit	KD [13]	73.33	72.98	73.54	74.92	70.66	73.08	70.67
	KD+Ours	76.62	74.36	74.37	76.11	71.43	74.17	71.48
	Δ	3.29	1.38	0.83	1.19	0.77	1.09	0.81
	KD+CTKD [24]	73.39	73.52	73.93	75.45	71.19	73.52	70.99
	KD+CTKD+Ours	76.67	74.47	74.58	76.08	71.34	74.01	71.39
	Δ	3.28	0.95	0.65	0.63	0.15	0.49	0.40
	DKD [57]	76.32	74.68	74.81	76.24	71.97	74.11	71.06
	DKD+Ours	77.01	74.81	74.89	76.39	<u>72.32</u>	<u>74.29</u>	71.85
	Δ	0.69	0.13	0.08	0.15	0.35	0.18	0.79
	MLKD [57]	77.08	<u>75.18</u>	<u>75.35</u>	<u>76.63</u>	72.19	74.11	<u>71.89</u>
	MLKD+Ours	**78.28**	**75.22**	**75.56**	**76.95**	**72.33**	**74.32**	**72.27**
	Δ	1.20	0.04	0.21	0.32	0.14	0.21	0.38

Heterogeneous

27 red 1 blue

Homogeneous

24 red 4 blue

김범준

[Logit Standardization in Knowledge Distillation]()

Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

CVPR 2024
(Highlight)

- **Experiment - 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 [53]	70.69	90.01	69.56	89.33
OFD [12]	70.81	89.98	71.25	90.34
CRD [39]	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 [57]	71.70	90.41	72.05	91.05
DKD+Ours	71.88 $_{+0.18}$	<u>90.58</u> $_{+0.17}$	72.85 $_{+0.80}$	91.23 $_{+0.18}$
MLKD [17]	<u>71.90</u>	90.55	<u>73.01</u>	<u>91.42</u>
MLKD+Ours	**72.08** $_{+0.18}$	**90.74** $_{+0.19}$	**73.22** $_{+0.21}$	**91.59** $_{+0.17}$

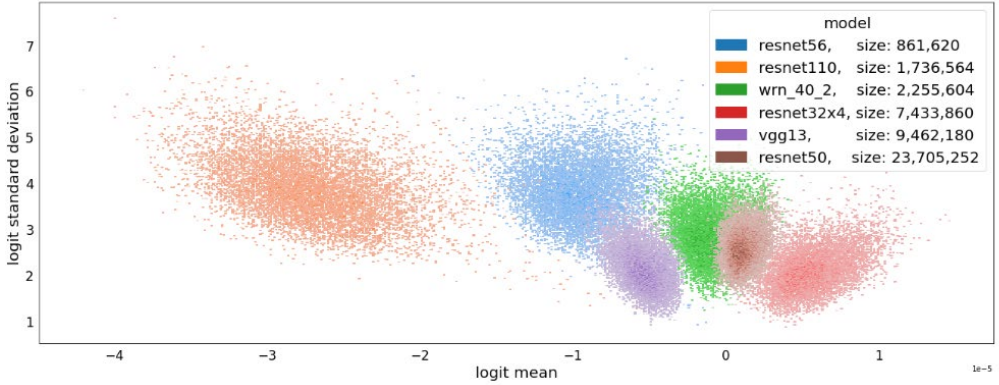
김범준

[Logit Standardization in Knowledge Distillation](#)

Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

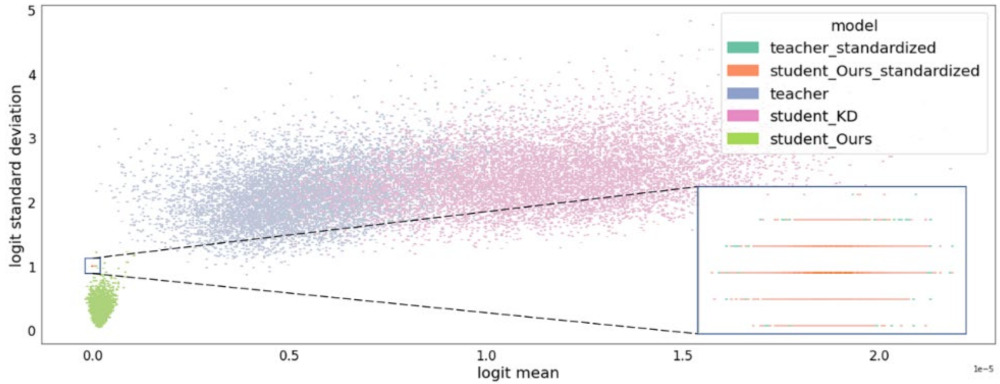
CVPR 2024
(Highlight)

- **Experiment - logit space 시각화**



(a) Teacher models of different sizes

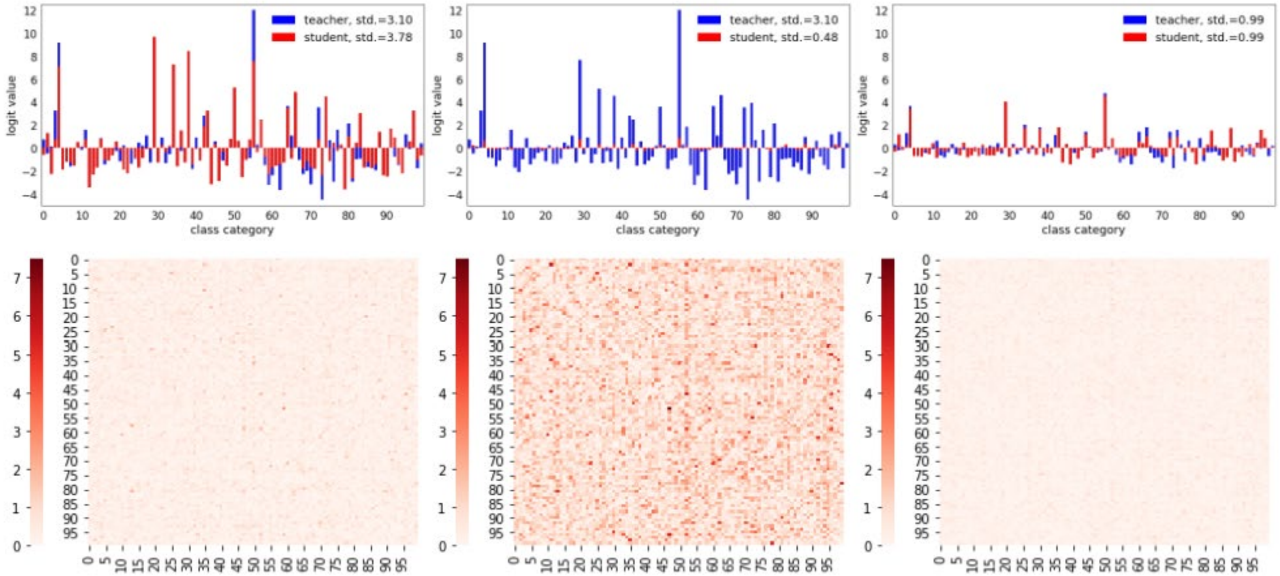
Teacher 모델에 따라서 Logit의 scale이 다르다
(parameter마다)



Standardized 했을때 엄청 넓게 퍼져있다가
엄청 좁은 영역에 일정하게 분포함

김범준

- **Experiment - 시각화**



(a) Vanilla KD
Mean: 0.27, Max: 3.03.

(b) Ours w/o $\mathcal{Z}$-score
Mean: 0.94, Max: 7.36.

(c) Ours w/ $\mathcal{Z}$-score
Mean: 0.18, Max:1.18.

평균만 빼면 일정한 격차로 유지

표준편차로 나눠야만 같은 스케일

Logit Standardization in Knowledge Distillation

Shangquan Sun1,2, Wenqi Ren3†, Jingzhi Li1, Rui Wang1,2, Xiaochun Cao3

CVPR 2024 (Highlight)

- **Experiment - robust**

Table 4. The ablation studies under different settings in our $\mathcal{Z}$-score. The base temperature τ is set to be 2. By default $\lambda_{\text{CE}} = 0.1$. The logit vector of teacher $\mathbf{v}_n$ and student $\mathbf{z}_n$ are abbreviated as $\mathbf{z}$ for succinctness. The teacher and student are ResNet32×4 and ResNet8×4.

λ_{KD}	$\mathbf{z}$ (KD)	$\mathbf{z} - \overline{\mathbf{z}}$	$\frac{\mathbf{z}}{\sigma(\mathbf{z})}$	$\frac{(\mathbf{z}-\overline{\mathbf{z}})}{\sigma(\mathbf{z})}$ (Ours)
0.9	73.60	73.37	73.79	74.14
3.0	74.38	74.33	75.86	76.11
6.0	74.45	74.82	76.44	76.56
9.0	73.33	73.94	76.30	76.62
12.0	68.29	71.56	76.49	76.56
15.0	65.34	62.01	76.42	76.61
18.0	63.45	61.31	76.18	76.33