Comparing Kullback-Leibler Divergence and Mean Squared Error Loss in Knowledge Distillation

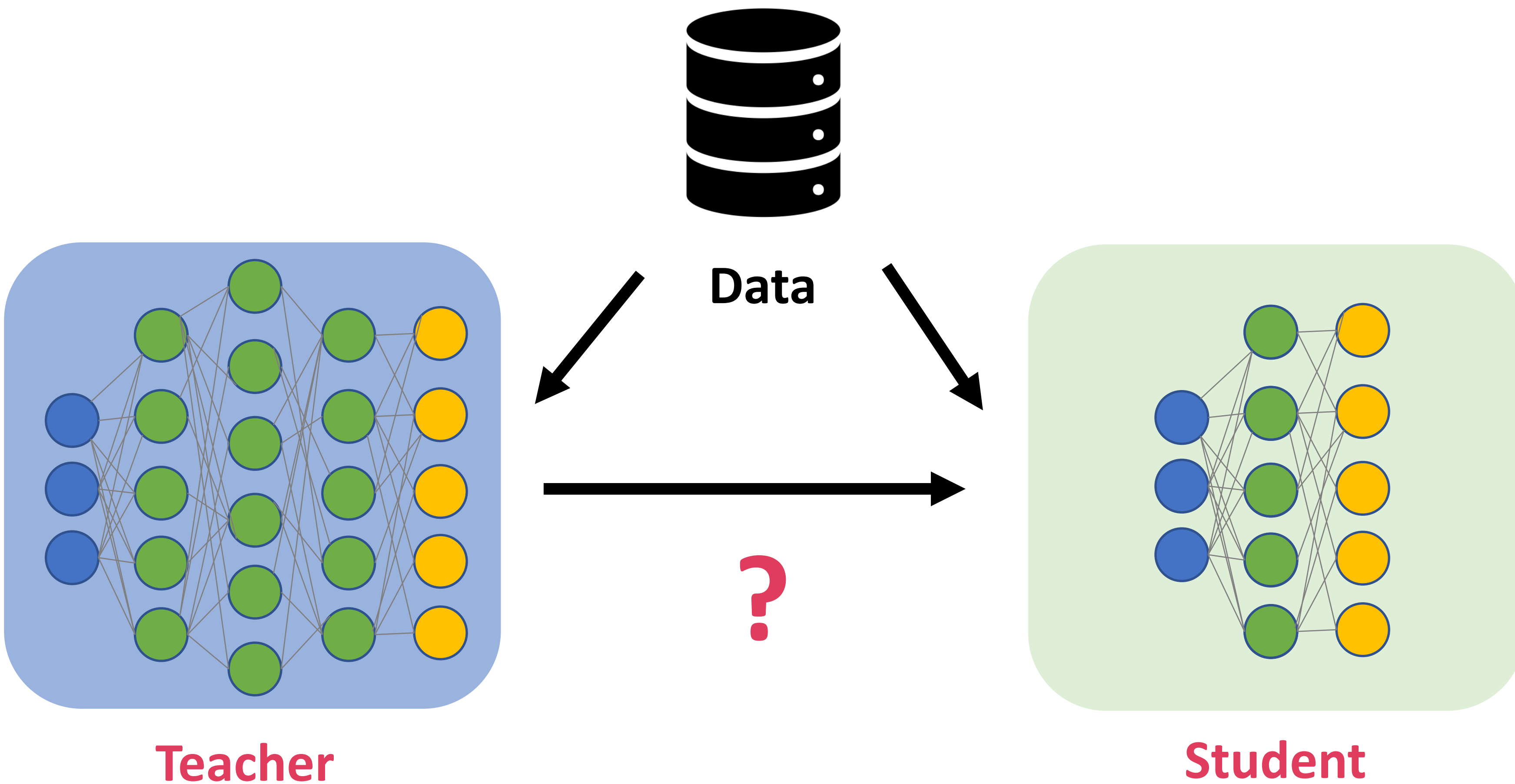
2021.06.04

Taehyeon Kim*
Jaehoon Oh*
Nakyil Kim
Sangwook Cho
Se-young Yun

Optimization and
Statistical Inference LAB

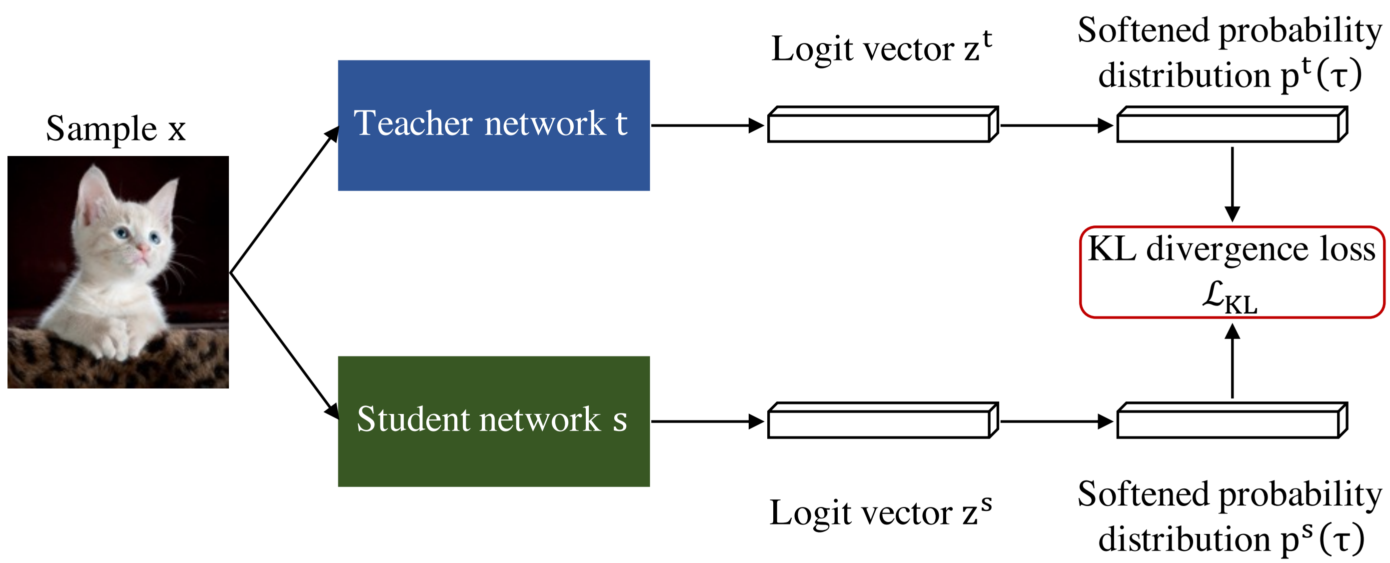
Why is Knowledge Distillation (KD) Beneficial

- One of the most potent model compression techniques.
- Knowledge is transferred from a cumbersome model (teacher) to a single small model (student).



Overview of KD

- KD has evolved to design a new objective function
  - KL divergence loss with temperature scaling



Degree of softness (temperature scaling)

- KD has evolved to design a new objective function
  - KL divergence loss with temperature scaling
  - Little Understanding of how the degree of softness affects the performance.

1. Learning from Ground Truth   2. Learning from Teacher model

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{CE}(\boldsymbol{p}^s(1), \boldsymbol{y}) + \alpha\mathcal{L}_{KL}(\boldsymbol{p}^s(\tau), \boldsymbol{p}^t(\tau)),$$

$$\mathcal{L}_{CE}(\boldsymbol{p}^s(1), \boldsymbol{y}) = \sum_j -\boldsymbol{y}_j \log \boldsymbol{p}_j^s(1)$$

$$\mathcal{L}_{KL}(\boldsymbol{p}^s(\tau), \boldsymbol{p}^t(\tau)) = \tau^2 \sum_j \boldsymbol{p}_j^t(\tau) \log \frac{\boldsymbol{p}_j^t(\tau)}{\boldsymbol{p}_j^s(\tau)}$$

$$\boxed{\boldsymbol{p}_k^f(\tau) = \frac{\exp(\boldsymbol{z}_k^f/\tau)}{\sum_{j=1}^K \exp(\boldsymbol{z}_j^f/\tau)}}$$

Preliminaries: respects to previous assumption

- Conventional assumption for KD

Derivative $$\frac{\partial \mathcal{L}_{KL}}{\partial z_k^s} \approx \tau \left(\frac{1 + z_k^s/\tau}{K + \sum_j z_j^s/\tau} - \frac{1 + z_k^t/\tau}{K + \sum_j z_j^t/\tau} \right)$$

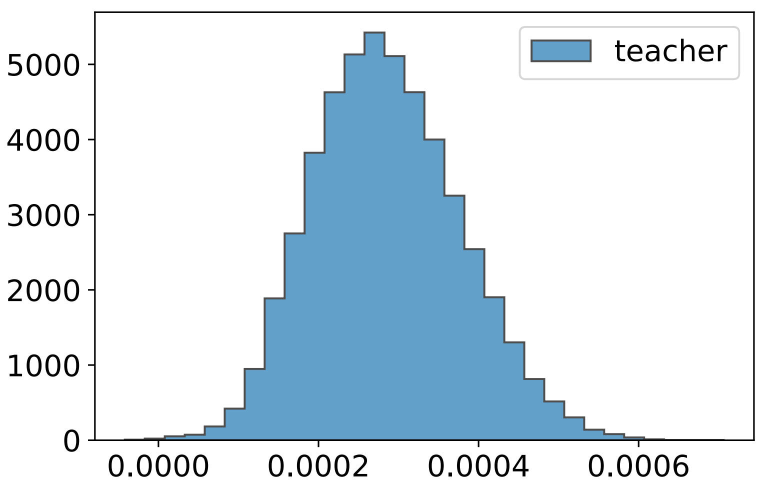
Assumption $\sum_j z_j^s = 0$ and $\sum_j z_j^t = 0$

Conclusion $$\frac{\partial \mathcal{L}_{KL}}{\partial z_k^s} \approx \frac{1}{K}(z_k^s - z_k^t)$$ then, $\mathcal{L}_{KL} = \mathcal{L}_{MSE}$????

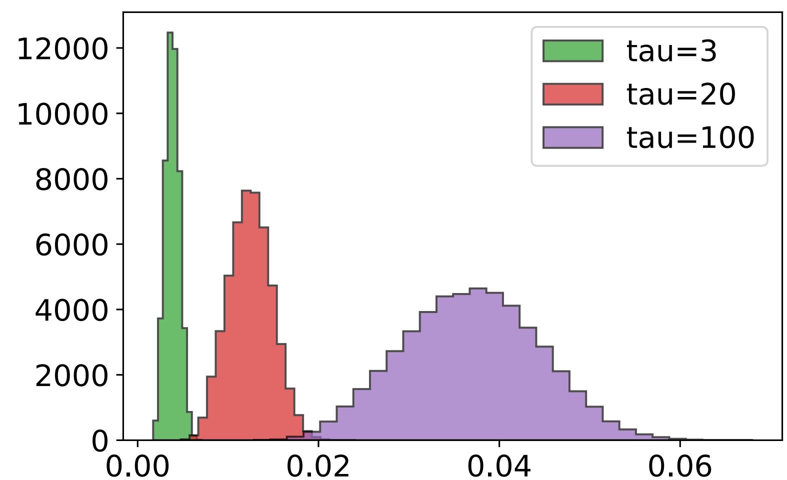
Preliminaries: respects to previous assumption

- Conventional assumption does not seem appropriate.

Assumption $\sum_j z_j^s = 0$ and $\sum_j z_j^t = 0$



$\neq$



$\neq \mathbf{0}$

Changes on τ

- Depending on τ, $\mathcal{L}_{KL}$ plays different roles
  - $\tau \to 0$: Label matching
  - $\tau \to \infty$: Logit matching

0 ⟷ ∞

$$\lim_{\tau \to 0} \frac{1}{\tau} \frac{\partial \mathcal{L}_{KL}}{\partial \boldsymbol{z}_k^s} = \mathbf{1}_{[\arg\max_j \boldsymbol{z}_j^s = k]} - \mathbf{1}_{[\arg\max_j \boldsymbol{z}_j^t = k]}$$

$$\begin{aligned}\lim_{\tau \to \infty} \frac{\partial \mathcal{L}_{KL}}{\partial \boldsymbol{z}_k^s} &= \frac{1}{K^2} \sum_{j=1}^{K} \left((\boldsymbol{z}_k^s - \boldsymbol{z}_j^s) - (\boldsymbol{z}_k^t - \boldsymbol{z}_j^t) \right) \\ &= \frac{1}{K} \left(\boldsymbol{z}_k^s - \boldsymbol{z}_k^t \right) - \frac{1}{K^2} \sum_{j=1}^{K} \left(\boldsymbol{z}_j^s - \boldsymbol{z}_j^t \right)\end{aligned}$$

Extension from KL loss to MSE Loss

- Some term is generated.
- *This term hinders complete logit matching (MSE loss) by shifting the mean of the elements in the logit.*

$$\lim_{\tau\to\infty} \frac{\partial \mathcal{L}_{KL}}{\partial \boldsymbol{z}_k^s} = \frac{1}{K^2}\sum_{j=1}^{K}\left((\boldsymbol{z}_k^s - \boldsymbol{z}_j^s) - (\boldsymbol{z}_k^t - \boldsymbol{z}_j^t)\right)$$

$$= \frac{1}{K}\left(\boldsymbol{z}_k^s - \boldsymbol{z}_k^t\right) - \frac{1}{K^2}\sum_{j=1}^{K}\left(\boldsymbol{z}_j^s - \boldsymbol{z}_j^t\right)$$

$$\downarrow$$

$$\lim_{\tau\to\infty} \nabla_{\boldsymbol{z}^s}\mathcal{L}_{KL} = \frac{1}{K}\left(\boldsymbol{z}^s - \boldsymbol{z}^t\right) - \frac{1}{K^2}\sum_{j=1}^{K}\left(\boldsymbol{z}_j^s - \boldsymbol{z}_j^t\right)\cdot \mathbb{1}$$

$$\downarrow$$

Bounded Convergence Theorem using each partial derivatives

$$\lim_{\tau\to\infty} \mathcal{L}_{KL} = \frac{1}{2K}||\boldsymbol{z}^s - \boldsymbol{z}^t||_2^2 + \delta_\infty = \frac{1}{2K}\mathcal{L}_{MSE} + \delta_\infty$$

$$\delta_\infty = -\frac{1}{2K^2}\left(\sum_{j=1}^{K}\boldsymbol{z}_j^s - \sum_{j=1}^{K}\boldsymbol{z}_j^t\right)^2 + Constant \longrightarrow$$

Analysis on this term!!

Theoretical Analysis on δ_∞

Optimization and Statistical Inference LAB

- Lower bound for δ_∞

$$\delta_\infty \approx -\frac{1}{2K^2}\left(\sum_{j=1}^{K} \boldsymbol{z}_j^s\right)^2 = -\frac{1}{2K^2}\left(\sum_{j=1}^{K}\sum_{n=1}^{d} W_{j,n}^s \boldsymbol{r}_n^s\right)^2$$

$$= -\frac{1}{2K^2}\left(\sum_{n=1}^{d} \boldsymbol{r}_n^s \sum_{j=1}^{K} W_{j,n}^s\right)^2$$

$$\geq -\frac{1}{2K^2}\left(\sum_{n=1}^{d}\left(\sum_{j=1}^{K} W_{j,n}^s\right)^2\right)\left(\sum_{n=1}^{d} \boldsymbol{r}_n^{s\,2}\right)$$

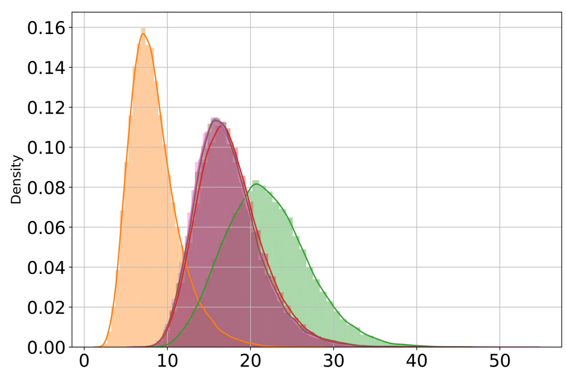
$$(\because \text{Cauchy-Schwartz inequality})$$

$$= -\frac{1}{2K^2}||\boldsymbol{r}^s||_2^2\left(\sum_{n=1}^{d}\left(\sum_{j=1}^{K} W_{j,n}^s\right)^2\right)$$

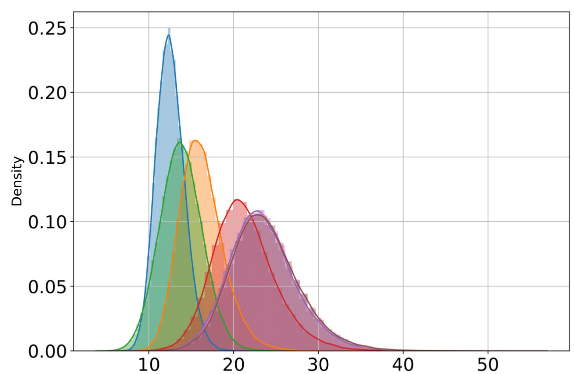
- Increasing the norm of r (pre-logit: input of fully-connected layer)
- De-shinkage effects on weight templates

Empirical Analysis on δ_∞

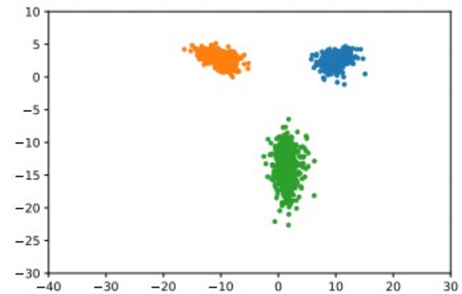
- Logit & Pre-logit behavior
- 2-D Projection Visualization



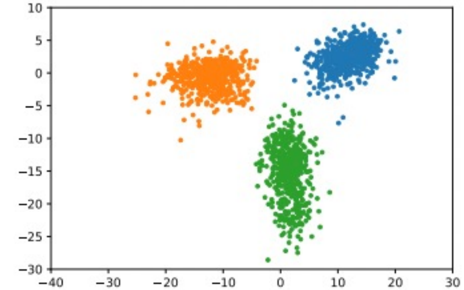
$||z^s - z^t||$



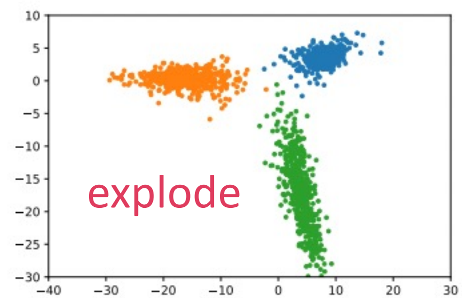
$||r||$



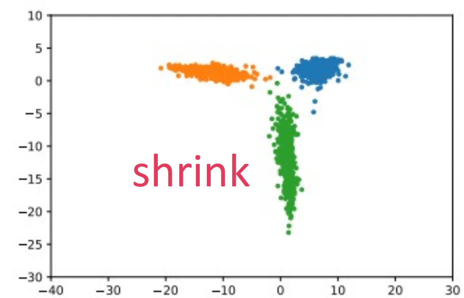
(a) $t, \mathcal{L}_{CE}$ (Train)



(b) $s, \mathcal{L}_{CE}$ (Train)



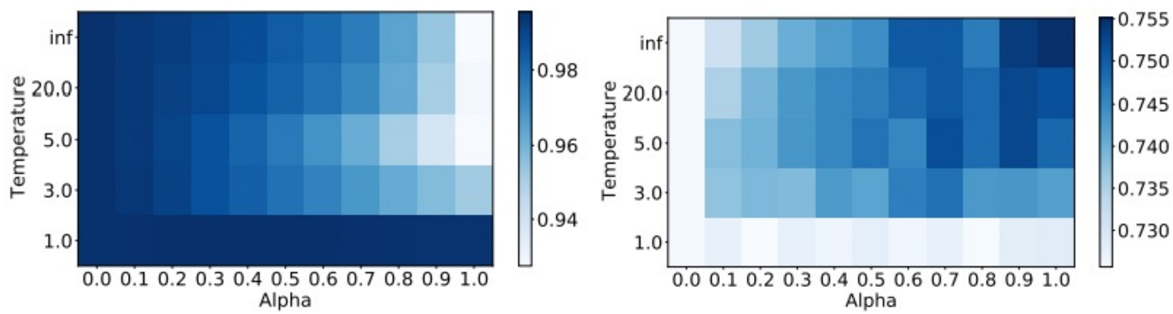
(c) $s, \mathcal{L}_{KL}(\tau = \infty)$ (Train)



(d) $s, \mathcal{L}_{MSE}$ (Train)

Empirical Results

- Training Accuracy and Test Accuracy (CIFAR-100)
- With perfectly trained teacher, **MSE is the best**!!



(a) Training accuracy. (b) Test accuracy.

Student	$\mathcal{L}_{CE}$	$\mathcal{L}_{KL}$ τ=1	$\mathcal{L}_{KL}$ τ=3	$\mathcal{L}_{KL}$ τ=5	$\mathcal{L}_{KL}$ τ=20	$\mathcal{L}_{KL}$ $\tau=\infty$	$\mathcal{L}_{MSE}$
WRN-16-2	72.68	72.90	74.24	74.88	75.15	75.51	**75.54**
WRN-16-4	77.28	76.93	78.76	78.65	78.84	78.61	**79.03**
WRN-28-2	75.12	74.88	76.47	76.60	**77.28**	76.86	**77.28**
WRN-28-4	78.88	78.01	78.84	79.36	79.72	79.61	**79.79**
WRN-40-6	79.11	79.69	79.94	79.87	79.82	79.80	**80.25**

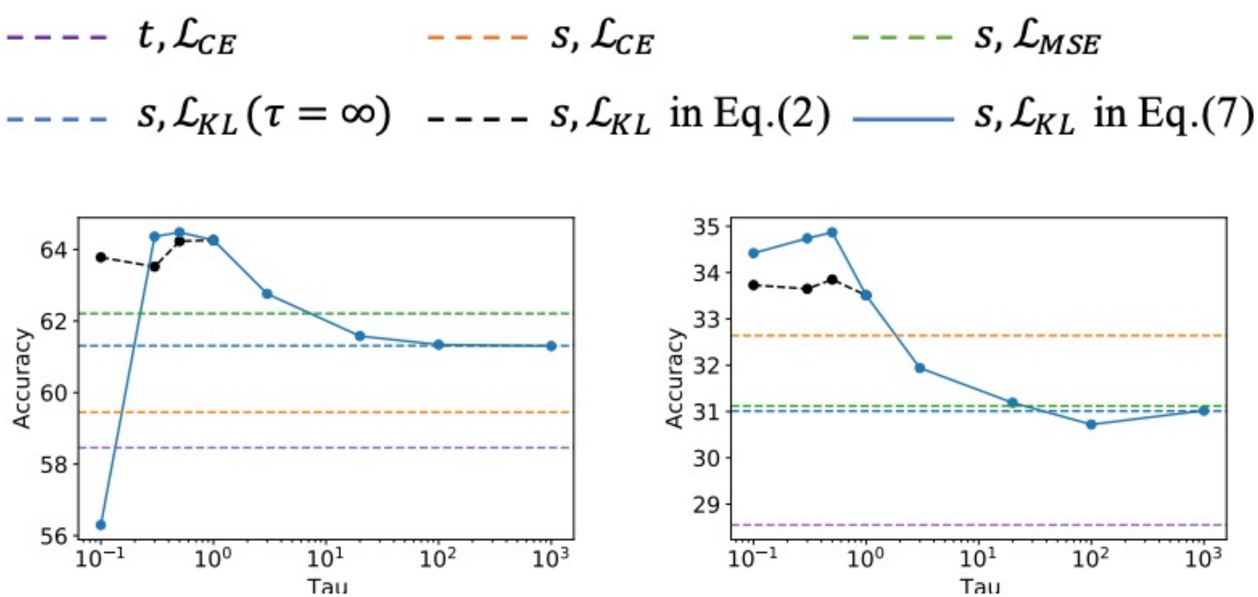
Empirical Results

- **MSE is also the best compared to other alternatives**!!

Student	Baseline	SKD [2015]	FitNets [2014]	AT [2016a]	Jacobian [2018]	FT [2018]	AB [2019b]	Overhaul [2019a]	MSE
WRN-16-2	72.68	73.53	73.70	73.44	73.29	74.09	73.98	**75.59**	75.54
WRN-16-4	77.28	78.31	78.15	77.93	77.82	78.28	78.64	78.20	**79.03**
WRN-28-2	75.12	76.57	76.06	76.20	76.30	76.59	76.81	76.71	**77.28**

Extreme small τ

- Robustness to Noisy Labels
- On severe noise rate (80%), low τ is better than others!
- It happens due to teacher's poor generalization.
- Label Matching is better!



(a) Symmetric noise 40% (b) Symmetric noise 80%

Adequate τ (Noisy Teacher)

- If your teacher model is not perfectly trained,
- But has the test accuracy around 80~90% then,
- the optimal solution for τ may be some number >1.

Student	$\mathcal{L}_{KL}$						$\mathcal{L}_{MSE}$
	τ=0.1	τ=0.5	τ=1	τ=5	τ=20	τ=∞	
WRN-16-2	51.64	52.07	51.36	50.11	49.69	49.46	49.20

Table 4: Top-1 test accuracies on CIFAR-100. WRN-28-4 is used as a teacher for $\mathcal{L}_{KL}$ and $\mathcal{L}_{MSE}$. Here, the teacher (WRN-28-4) was not fully trained. The training accuracy of the teacher network is 53.77%.

Student	$\mathcal{L}_{CE}$	$\mathcal{L}_{KL}$ (Standard)	$\mathcal{L}_{KL}$ (τ=20)	$\mathcal{L}_{MSE}$
ResNet-50	76.28	77.15	77.52	75.84

Table 5: Test accuracy on the ImageNet dataset. We used a (teacher, student) pair of (ResNet-152, ResNet-50). We include the results of the baseline and $\mathcal{L}_{KL}$ (standard) from [Heo *et al.*, 2019a]. The training accuracy of the teacher network is 81.16%.

Sequential Distillation

- According to the changes of objective functions,
- The performance varies significantly even under the usage of the same architectures.

WRN-28-4	WRN-16-4	WRN-16-2	Test accuracy
X	X	$\mathcal{L}_{CE}$	72.68 %
X	$\mathcal{L}_{CE}$ (77.28%)	$\mathcal{L}_{KL}(\tau = 3)$	74.84 %
		$\mathcal{L}_{KL}(\tau = 20)$	75.42 %
		$\mathcal{L}_{MSE}$	75.58 %
$\mathcal{L}_{CE}$ (78.88%)	X	$\mathcal{L}_{KL}(\tau = 3)$	74.24 %
		$\mathcal{L}_{KL}(\tau = 20)$	75.15 %
		$\mathcal{L}_{MSE}$	75.54 %
$\mathcal{L}_{CE}$ (78.88%)	$\mathcal{L}_{KL}(\tau = 3)$ (78.76%)	$\mathcal{L}_{KL}(\tau = 3)$	74.52 %
		$\mathcal{L}_{KL}(\tau = 20)$	75.47 %
		$\mathcal{L}_{MSE}$	**75.78 %**
	$\mathcal{L}_{MSE}$ (79.03%)	$\mathcal{L}_{KL}(\tau = 3)$	74.83 %
		$\mathcal{L}_{KL}(\tau = 20)$	75.47 %
		$\mathcal{L}_{MSE}$	75.48 %

E.O.D.

Optimization and Statistical Inference LAB
potter32@kaist.ac.kr
www.osi.kaist.ac.kr