Asymmetric Temperature Scaling Makes Larger Networks Teach Well Again

Borui Zhao[1] Quan Cui[2] Renjie Song[1] Yiyu Qiu[1,3] Jiajun Liang[1]

[1]MEGVII Technology [2]Waseda University [3]Tsinghua University

zhaoborui.gm@gmail.com, cui-quan@toki.waseda.jp,

chouyy18@mails.tsinghua.edu.cn, {songrenjie, liangjiajun}@megvii.com

Presenter: Seonghak KIM

Introduction

- **Vanilla KD**
  - Kullback-Leibler (KL) divergence between output probabilities, $\mathrm{KL}\left(\boldsymbol{p}^{\mathcal{T}} \middle\| \boldsymbol{p}^{\mathcal{S}}\right)$

  ➔ ***more accurate teacher don't necessarily teach better.***
- **Questions**
  - ***What's the reason that more complex teachers can't teach well?***
    - Decomposition of teacher's probability
      - Correct Guidance: correct class's probability
      - Smooth Regularization: average probability of wrong classes (DA)
      - Class Discriminability: variance of wrong class probabilities (DV)
    - Complex teacher are *over-confident*. (***larger score for correct / less varied score for the wrong classes***)

    ∴ Uniform temperature → effective class discriminability ↓ (distinctness of wrong class probability ↓)
  - ***Is it impossible to make larger teachers teach better through simple operations (temperature scaling)?***
    - *Asymmetric Temperature Scaling (ATS)*: separate higher/lower temperature for the correct/wrong logit instead of uniform temperature → variance of wrong class probabilities (DV) ↑ (discriminability ↑)

Background

- **Notations**
  - Input $\mathbf{x} \rightarrow$ logits $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^C \rightarrow$ softened probability $\boldsymbol{p}_c(\tau) = \frac{\exp(\mathbf{f}_c(\mathbf{x})/\tau)}{\sum_{j=1}^{C}\exp(\mathbf{f}_j(\mathbf{x})/\tau)}$
  - Correct logit $= \mathbf{f}_y$, Correct probability $= \boldsymbol{p}_y$
  - Wrong logits $= \mathbf{g}\left(= [\mathbf{f}_c]_{c\neq y}\right)$, Wrong probability $= \mathbf{q}\left(= [\boldsymbol{p}_c]_{c\neq y}\right)$
  - $\tilde{\mathbf{q}}_{c\prime} = \frac{\exp(\mathbf{g}_{c\prime}(\mathbf{x})/\tau)}{\sum_{j=1}^{C}\exp(\mathbf{g}_j(\mathbf{x})/\tau)}$; for only the wrong logits
    - $\tilde{\mathbf{q}} \neq \mathbf{q} \because \sum_{c\prime}\tilde{\mathbf{q}}_{c\prime} = 1$ and $\sum_{c\neq y}\boldsymbol{p}_c = 1 - \boldsymbol{p}_y$

- **Note**
  - **The effectiveness is more related to the *distinctness between wrong classes* rather than all classes.**
    - The variance of wrong class probabilities is focused instead of all classes.
    - $\max(\text{var}_{\text{all}}) \neq \max(\text{var}_{\text{wrong}})$

Decomposition

- **Distillation loss**

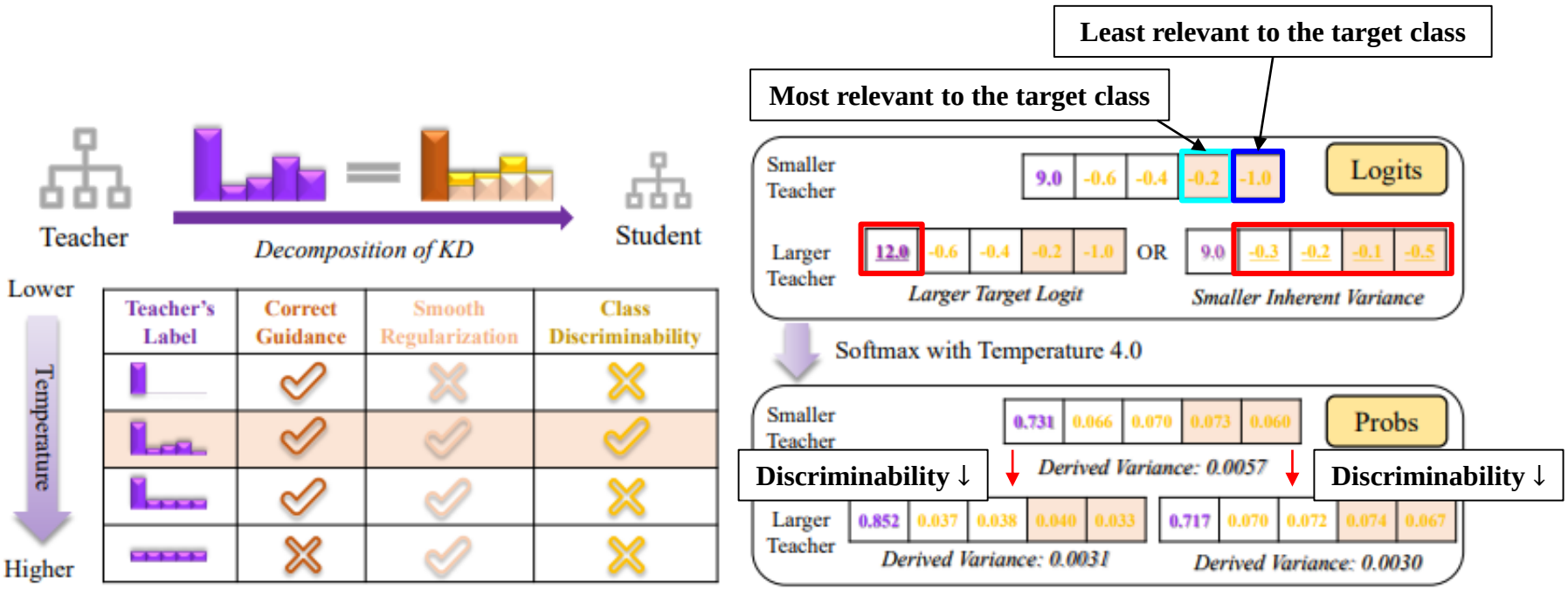
$$\mathcal{L}_{\mathrm{KD}} = -\lambda\tau^2 \sum_{c=1}^{C} \boldsymbol{p}_c^{\mathcal{T}} \log\left(\boldsymbol{p}_c^{\mathcal{S}}\right)$$

$$= -\lambda\tau^2 \left(\boldsymbol{p}_y^{\mathcal{T}} \log\left(\boldsymbol{p}_y^{\mathcal{S}}\right) - \sum_{c\neq y}^{C} e\left(\mathbf{q}^{\mathcal{T}}\right) \log \boldsymbol{p}_c^{\mathcal{S}} - \sum_{c\neq y}^{C} \left(\boldsymbol{p}_c^{\mathcal{T}} - e\left(\mathbf{q}^{\mathcal{T}}\right) \right) \log \boldsymbol{p}_c^{\mathcal{S}} \right)$$

  - $\boldsymbol{p}_y^{\mathcal{T}} \log\left(\boldsymbol{p}_y^{\mathcal{S}}\right)$: Correct Guidance guarantees *correctness*
  - $\sum_{c\neq y}^{C} e\left(\mathbf{q}^{\mathcal{T}}\right) \log \boldsymbol{p}_c^{\mathcal{S}}$: Smooth Regularization
    - $e\left(\mathbf{q}^{\mathcal{T}}\right) = \frac{1}{C-1} \sum_{c\neq y} \boldsymbol{p}_c$: average of wrong class probability (DA)
  - $\sum_{c\neq y}^{C} \left(\boldsymbol{p}_c^{\mathcal{T}} - e\left(\mathbf{q}^{\mathcal{T}}\right) \right) \log \boldsymbol{p}_c^{\mathcal{S}}$: Class Discriminability (which classes are more related to the correct class?)
    - $v(\mathbf{q}) = \frac{1}{C-1} \sum_{c\neq y} \left(\boldsymbol{p}_c^{\mathcal{T}} - e\left(\mathbf{q}^{\mathcal{T}}\right) \right)$: variance of wrong class probability (DV)
    - cf. $v(\tilde{\mathbf{q}})$: *Inherent Variance (IV)* because it only depends on wrong classes' logits.

Decomposition

- **Figure 1**
  - **Left**
    - Temperature ↑ ➔ Correct guidance ↓, smooth regularization ↑, and class discriminability ↑→↓
      - Too high or too low temperature leads to smaller class discriminability (less distinctness among wrong classes).
  - **Right**
    - Larger teachers are over-confident (larger target logit or smaller inherent variance).



[Fig. 1]

Theoretical Analysis

- **Lemma 1. (Variance of Softened Probabilities)**
  - $v(\boldsymbol{p})$ (of all samples) monotonically ↓ as τ ↑.
    - Temperature ↑ → probability distribution become flat (uniform). (less variance for average)
- **Assumption 2.**
  - The target logits is higher than other classes' logits. $(\mathbf{f}_y \geq \mathbf{f}_c)$
- **Proposition 3.**
  - Under Assumption 2, τ ↑ → $\boldsymbol{p}_y$ ↓ and $e(\mathbf{q})$ ↑, monotonically.
    - Due to flatten distribution, target probability value ↓ and others probability value ↑
    - ∴ higher DA and strengthen the smooth regularization term
- **Proposition 4. (DV vs IV)**
  - $v(\mathbf{q}) = (C-1)^2 e^2(\mathbf{q}) v(\tilde{\mathbf{q}})$
    - τ ↑→ $e(\mathbf{q})$ ↑ **(**Prop. 3**)** , $v(\tilde{\mathbf{q}})$ ↓ (Lemma. 1**)** → difficult to judge monotonicity of $v(\mathbf{q})$
    - Empirically, $v(\mathbf{q})$ ↑ → ↓ ∴ class discriminability ↑ → ↓

Theoretical Analysis

- **Remark 5.**
  - Fixing $\mathbf{g}$ and τ, higher $\mathbf{f}_y \rightarrow$ higher $\boldsymbol{p}_y$ (smaller $e(\mathbf{q})$)
    - $\because$ target probability $\uparrow \rightarrow$ others probability $\downarrow \rightarrow e(\mathbf{q}) \downarrow$
- **Remark 6.**
  - Fixing τ, less varied wrong logits $\mathbf{g}$ → less varied $\tilde{\mathbf{q}}$ (smaller $v(\tilde{\mathbf{q}})$)
- **Corollary 7. ($\mathcal{T}_1$ is larger teacher than $\mathcal{T}_2$)**
  - If $\mathbf{f}_y^{\mathcal{T}_1} \geq \mathbf{f}_y^{\mathcal{T}_2}$ and $\mathbf{g}^{\mathcal{T}_1} \approx \mathbf{g}^{\mathcal{T}_2}$, then $\boldsymbol{p}_y^{\mathcal{T}_1} \geq \boldsymbol{p}_y^{\mathcal{T}_2}$ (Rema. 5) and $v(\tilde{\mathbf{q}}^{\mathcal{T}_1}) \approx v(\tilde{\mathbf{q}}^{\mathcal{T}_2})$. Hence, $v(\mathbf{q}^{\mathcal{T}_1}) \leq v(\mathbf{q}^{\mathcal{T}_2})$.
    - According to $\boldsymbol{p}_y^{\mathcal{T}_1} \geq \boldsymbol{p}_y^{\mathcal{T}_2} \rightarrow e(\mathbf{q}^{\mathcal{T}_1}) \leq e(\mathbf{q}^{\mathcal{T}_2})$ and Prop. 4.
  - If $\mathbf{f}_y^{\mathcal{T}_1} \approx \mathbf{f}_y^{\mathcal{T}_2}$ and $v(\mathbf{g}^{\mathcal{T}_1}) \leq v(\mathbf{g}^{\mathcal{T}_2})$, then $\boldsymbol{p}_y^{\mathcal{T}_1} \approx \boldsymbol{p}_y^{\mathcal{T}_2}$ and $v(\tilde{\mathbf{q}}^{\mathcal{T}_1}) \leq v(\tilde{\mathbf{q}}^{\mathcal{T}_2})$ (Rema. 6). Hence, $v(\mathbf{q}^{\mathcal{T}_1}) \leq v(\mathbf{q}^{\mathcal{T}_2})$.
    - According to $\boldsymbol{p}_y^{\mathcal{T}_1} \approx \boldsymbol{p}_y^{\mathcal{T}_2} \rightarrow e(\mathbf{q}^{\mathcal{T}_1}) \approx e(\mathbf{q}^{\mathcal{T}_2})$ and Prop. 4.

$\therefore$ larger teacher tend to be over-confident (larger target logit $\mathbf{f}_y$ or smaller variance of wrong logits $v(\mathbf{g})$) → smaller derived variance $v(\mathbf{q})$ (class discriminability $\downarrow$)

Asymmetric Temperature Scaling

- **Different temperatures to the logits of correct and wrong classes**
  - $\boldsymbol{p}_c(\tau_1, \tau_2) = \frac{\exp(\mathbf{f}_c/\tau_c)}{\sum_j^C \exp(\mathbf{f}_j/\tau_i)}, \tau_i = \mathcal{L}\{i = y\}\tau_1 + \mathcal{L}\{i \neq y\}\tau_2$ where $(\tau_1 > \tau_2 > 0)$
    - If the teacher outputs a larger target logits $\mathbf{f}_y$, a relatively larger τ_1 decrease $\mathbf{f}_y$ to a reasonable magnitude. (i.e., $\boldsymbol{p}_y \downarrow,\ e(\mathbf{q}) \uparrow \rightarrow \therefore v(\mathbf{q}) \uparrow$)
    - If the teacher outputs less varied logits $\mathbf{g}$, a relatively smaller τ_2 make $\mathbf{g}$ more diverse. (i.e., $v(\tilde{\mathbf{q}}) \uparrow \rightarrow v(\mathbf{q}) \uparrow$)

➔ **ATS make the distribution over wrong classes more discriminative.**

Experiments

- **Observations**
  - **Class discriminability matter in KD and correlates with the KD improvement.**
    - Without class discriminability → small and larger teachers teach worse significantly. (Fig. 2)
    - Fig. 3 shows that the teachers with a larger DV tend to guide better.
      - Dots in figure indicate that improvement is higher than 2%.

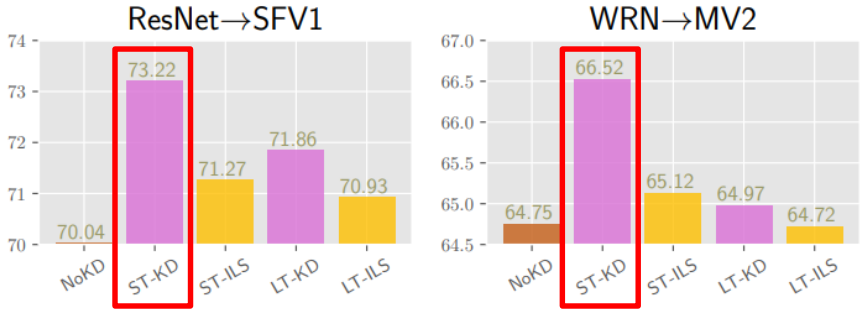


Figure 2: Student's test accuracies without KD ("NoKD"), with KD ("-KD"), and only with the first two terms in Eq. 2 ("-ILS"). "ST"/"LT" refers to "small/large teacher".

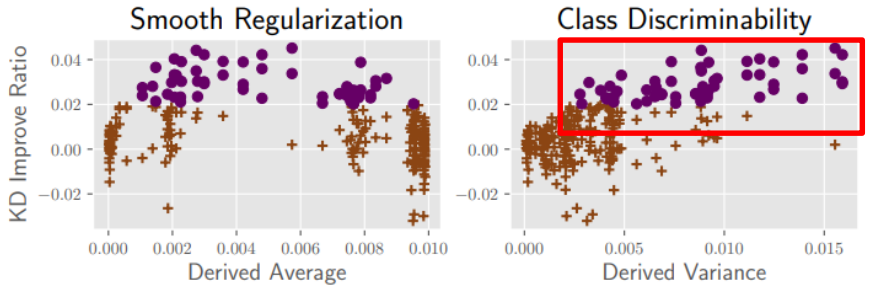


Figure 3: Correlations of *smooth regularization* (measured by *derived average*) and *class discriminability* (measured by *derived variance*) w.r.t. KD improvement ratio.

Experiments

- **Observations**
  - **Larger teachers provide a larger target logit or less varied wrong logits.**
    - On CIFAR-100, ResNet110 tend to generate a larger target logit ($\mathbb{E}_{\mathbf{x}}(\mathbf{f}_y) \approx 15.0$) than ResNet14 ($\mathbb{E}_{\mathbf{x}}(\mathbf{f}_y) \approx 10.0$)
    - On CIFAR-10, although $\mathbf{f}_y$ by WRN28-8 $\geq \mathbf{f}_y$ by WRN28-1, WRN28-8 gives smaller variance.

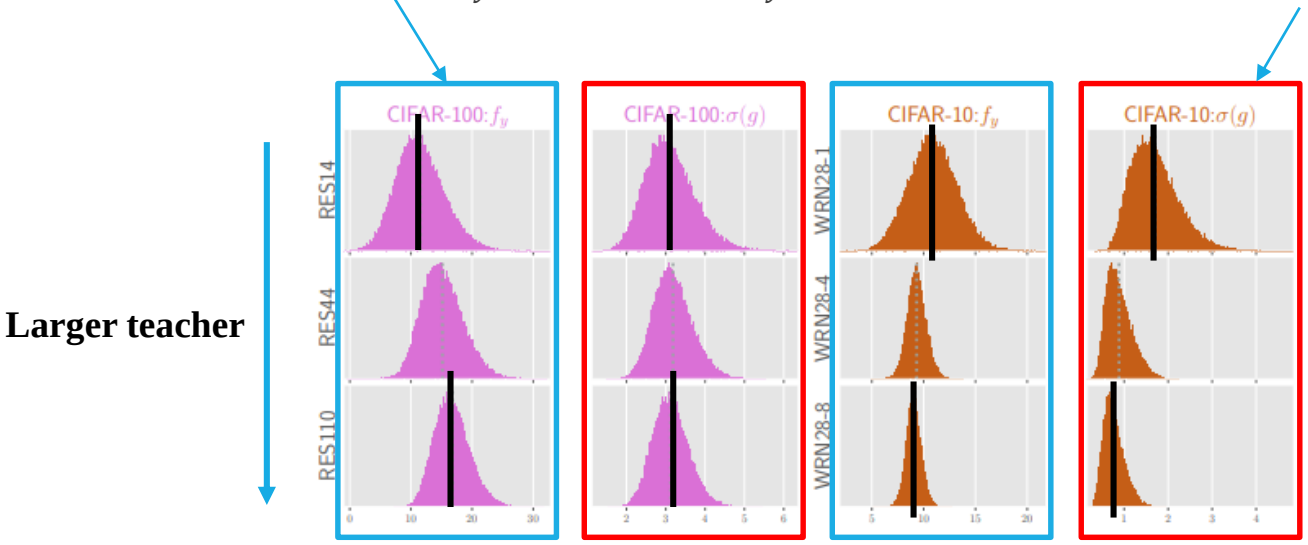


Figure 4: The distributions of the *target logit* ($\mathbf{f}_y$) and the *standard deviation of wrong logits* ($\sigma(\mathbf{g})$) of the 50K training samples on CIFAR-10/100. Rows show networks with various capacity.

Experiments

- **Observations**
  - **ATS could enlarge the derived variance of larger teachers.**
    - With ATS, larger teacher enhance DV while that teacher under traditional scaling experiences lower DV.
    - $\tau \uparrow \rightarrow e(\mathbf{q}) \uparrow$:smooth regularization ↑ (Prop. 3) (nearly same between various capacities)
    - $v(\mathbf{q})$ first increase and then decreases. (maximal of larger teachers' DV is smaller.)

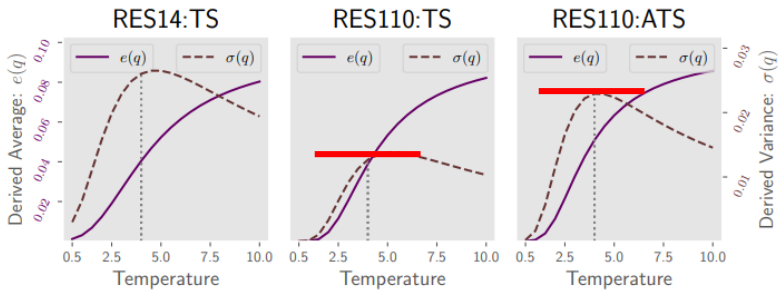


Figure 7: The change of *derived average* ($e(\mathbf{q})$) and *derived variance* ($v(\mathbf{q})$) as τ increases from 0.1 to 10.0 on CIFAR-10. The third one shows the results of ResNet110 with the proposed ATS. *DV* under TS is limited while ATS enlarges it.

Traditional scaling (top: small teacher, bottom: larger teacher)

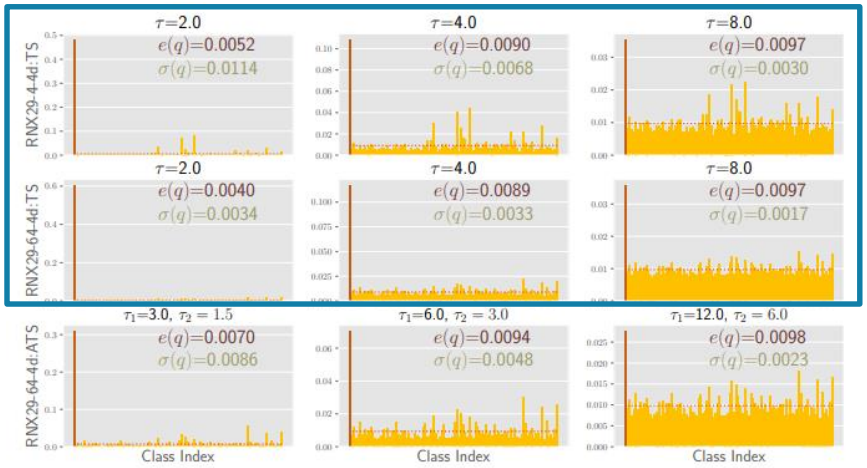


Figure 8: Probability vector visualization of a randomly selected training sample from CIFAR-100. The target class is $y = 1$. The bottom row shows applying ATS to the larger teacher.

Experiments

- **Performances**
  - **Using ATS, larger teachers teach well or better (↔ traditional scaling, TS).**

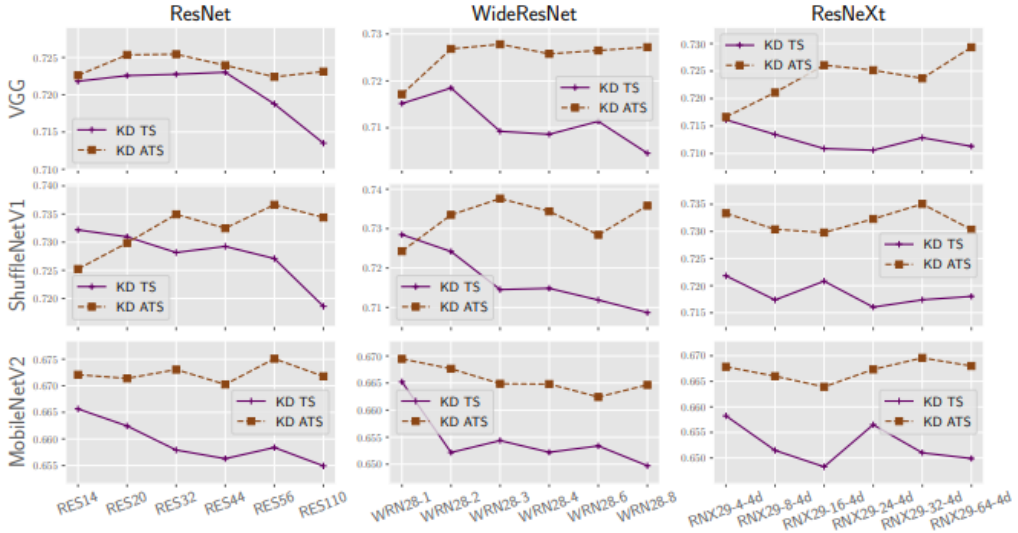


Figure 9: Distillation results via TS (solid curves) and ATS (dashed curves) on CIFAR-100. The x-axis of each figure shows teachers with various capacities.

Experiments

Performances

Comparisons with SOTA methods

- NoKD (w/o KD), ST-KD (guidance of smaller teacher), KD (guidance of larger teacher)

Table 2: Comparisons with SOTA methods on CIFAR-100. ResNet110, WRN28-8, and RNX29-64-4d are teachers. VGG8, SFV1, and MV2 are students. The area in gray shows the results of the ensemble. "KD+ATS" and "KD+ATS+Ens" are our methods.

Teacher	ResNet110 (74.09)			WRN28-8 (79.73)			RNX29-64-4d (79.91)			Avg
Student	VGG8	SFV1	MV2	VGG8	SFV1	MV2	VGG8	SFV1	MV2	
NoKD	69.92	70.04	64.75	69.92	70.04	64.75	69.92	70.04	64.75	68.24
ST-KD	72.30	73.22	66.56	71.85	72.85	**66.52**	71.61	72.18	65.82	70.32
KD	71.35	71.86	65.49	70.46	70.87	64.97	71.13	71.80	64.99	69.21
ESKD	71.88	72.02	65.92	71.13	71.32	65.09	71.09	71.27	64.83	69.39
TAKD	**72.71**	72.86	66.98	71.20	71.62	65.11	71.46	71.44	65.36	69.86
SCKD	70.38	70.61	64.59	70.83	70.52	65.19	70.33	70.92	64.86	68.69
KD+ATS	72.31	**73.44**	**67.18**	**72.72**	**73.58**	66.47	**72.93**	**73.03**	**66.80**	**70.94**
Ens	72.77	73.61	67.76	72.77	73.61	67.76	72.77	73.61	67.76	71.38
ResKD	73.89	**76.03**	69.00	73.84	**75.14**	67.69	74.64	75.43	68.10	72.64
KD+ATS+Ens	**74.86**	75.05	**69.50**	**74.60**	75.04	**68.79**	**75.34**	**75.47**	**69.82**	**73.16**

Experiments

- **Performances**
  - **Comparisons with SOTA methods**
    - NoKD (w/o KD), ST-KD (guidance of smaller teacher), KD (guidance of larger teacher)

Table 3: Comparisons with SOTA methods on TinyImageNet, CUB, and Stanford Dogs. WRN50-2 and RNX101-32-8d are teachers. AlexNet, SFV2, and MV2 are students.

	TinyImageNet			CUB			Stanford Dogs			Avg
Teacher	WRN50-2 (66.28)			RNX101-32-8d (79.50)			RNX101-32-8d (73.98)			
Student	ANet	SFV2	MV2	ANet	SFV2	MV2	ANet	SFV2	MV2	
NoKD	34.62	45.79	52.03	55.66	71.24	74.49	50.20	68.72	68.67	57.94
ST-KD	36.16	49.59	52.93	56.39	72.15	76.80	51.95	69.92	72.06	59.77
KD	35.83	48.48	52.33	55.10	71.89	76.45	50.22	68.48	71.25	58.89
ESKD	34.97	48.34	52.15	55.64	72.15	76.87	50.39	69.02	71.56	59.01
TAKD	36.20	48.71	52.44	54.82	71.53	76.25	50.36	68.94	70.61	58.87
SCKD	36.16	48.76	51.83	56.78	71.99	75.13	51.78	68.80	70.13	59.04
KD+ATS	**37.42**	**50.03**	**54.11**	**58.32**	**73.15**	**77.83**	**52.96**	**70.92**	**73.16**	**60.88**
Ens	39.37	50.69	56.40	59.84	74.43	77.47	54.04	71.65	72.53	61.82
ResKD	38.66	51.93	57.32	**62.60**	75.29	76.27	54.68	70.73	72.85	62.26
KD+ATS+Ens	**40.42**	**52.14**	**58.47**	62.00	**76.26**	**78.97**	**55.69**	**73.22**	**74.67**	**63.54**

Experiments

- **Performances**
  - **Ablation studies**
    - Setting $\tau_1 > \tau_2$ is better, especially, the setting of $\tau_2 \in [\tau_1 - 2, \tau_1 - 1]$ is recommended.
    - ATS improves the performances under various λ.

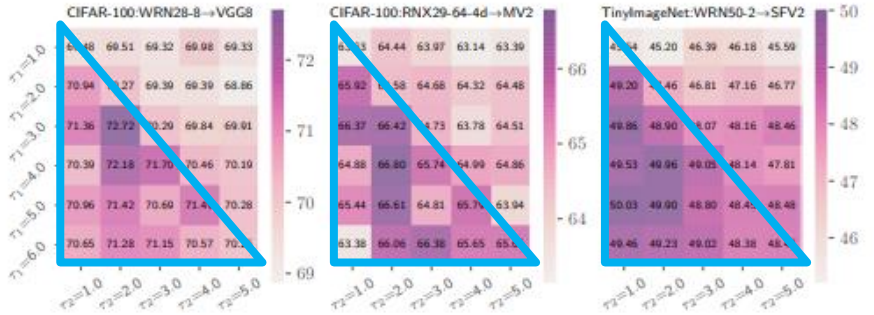


Figure 10: Ablation studies on asymmetric temperatures on CIFAR-100 and TinyImageNet (τ_1, τ_2 in Eq. 5).

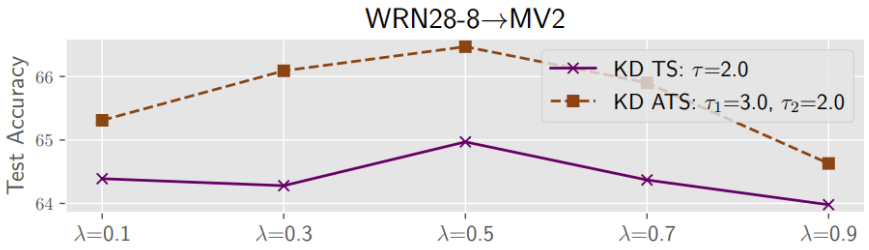


Figure 11: Ablation studies on the weighting of KD loss and CE loss on CIFAR-100 (λ in Eq. 1).

Conclusion

- **Decomposition into *correct guidance*, *smooth regularization*, and *class discriminability*.**
- **Over-confidence teachers can't utilize the *class discriminability* under TS.**
- ***Asymmetric Temperature Scaling (ATS)* to enhance the *DV* of larger teachers, making *more discriminative*, was proposed.**

Thank you.