# Asymmetric Temperature Scaling Makes Larger Networks Teach Well Again

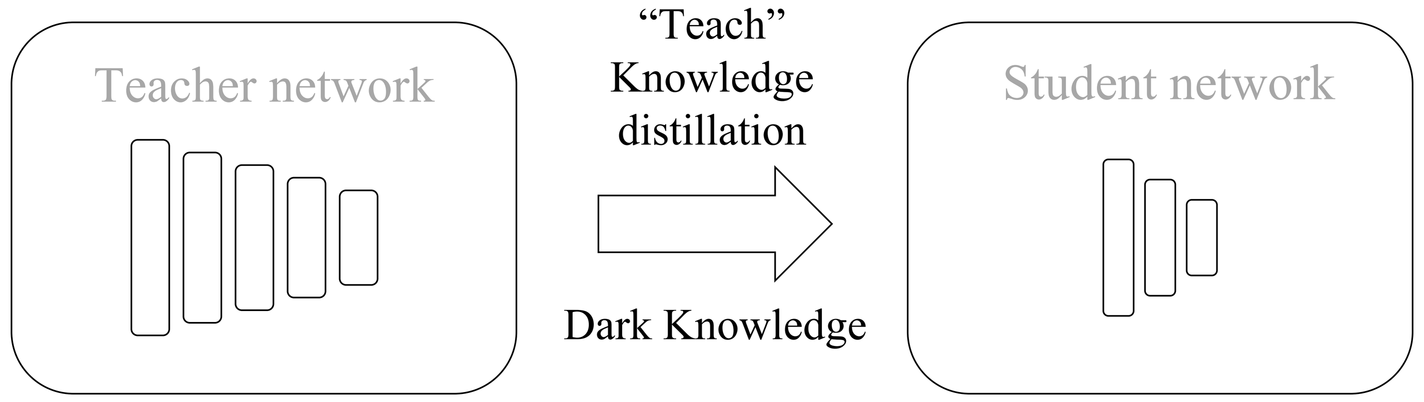
Xin-Chun Li, Wen-Shu Fan, De-Chuan Zhan
LAMDA Group, State Key Lab for Novel Software Technology, Nanjing University, China

Shaoming Song, Yinchuan Li, Bingshuai Li, Yunfeng Shao
Huawei Noah's Ark Lab

Email: lixc@lamda.nju.edu.cn, zhandc@nju.edu.cn Code: https://github.com/lxcnju/ATS-LargeKD

## Background

Knowledge Distillation can transfer the "knowledge" of large models to lightweight models:

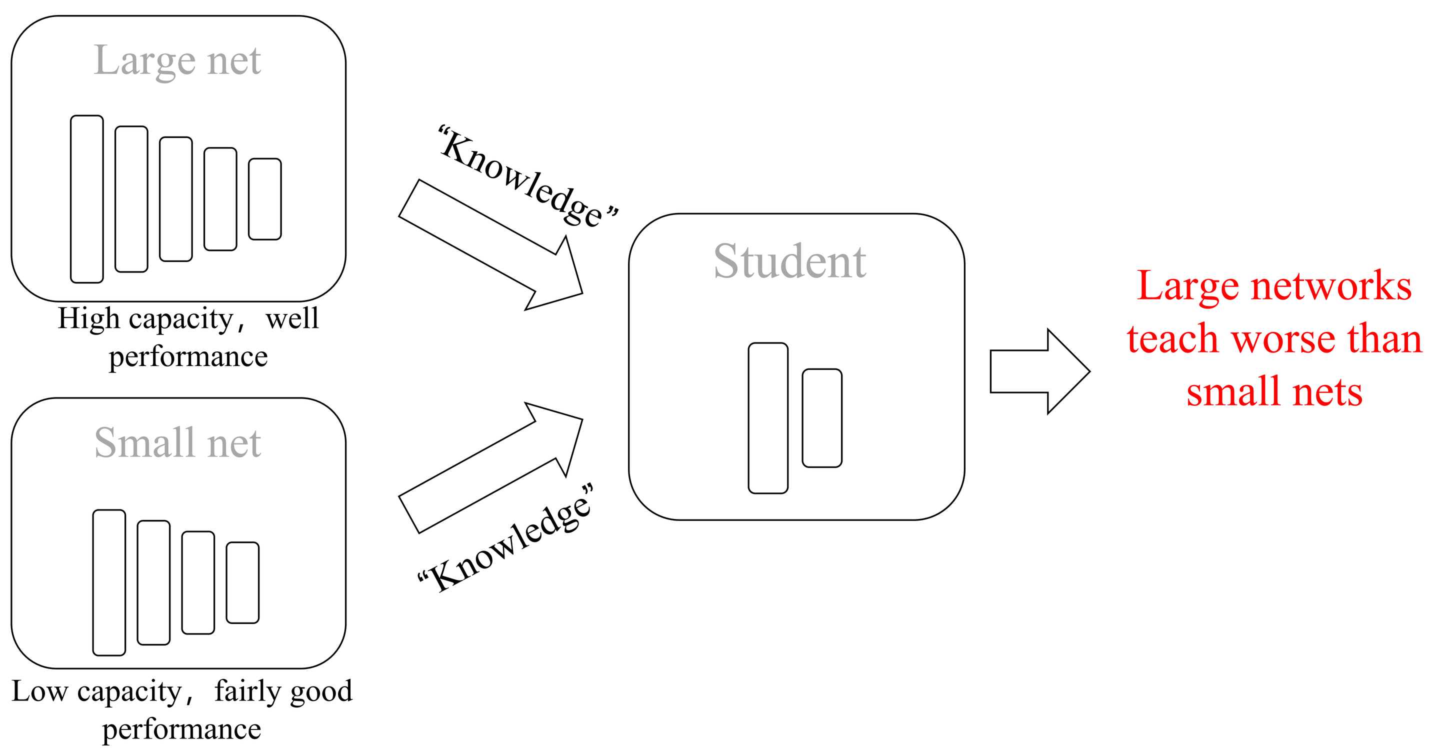



**Loss of student** = loss of classification + loss of KD

$$\ell = \underbrace{-(1-\lambda)\log \mathbf{p}_y^S(1)}_{\text{CE Loss}} \underbrace{-\lambda\tau^2 \sum_{c=1}^{C} \mathbf{p}_c^T(\tau)\log \mathbf{p}_c^S(\tau)}_{\text{KD Loss}},$$

$\mathbf{p}$: probs by network; $T, S$: teacher, student; $\tau$: temperature; $\lambda$: balance factor

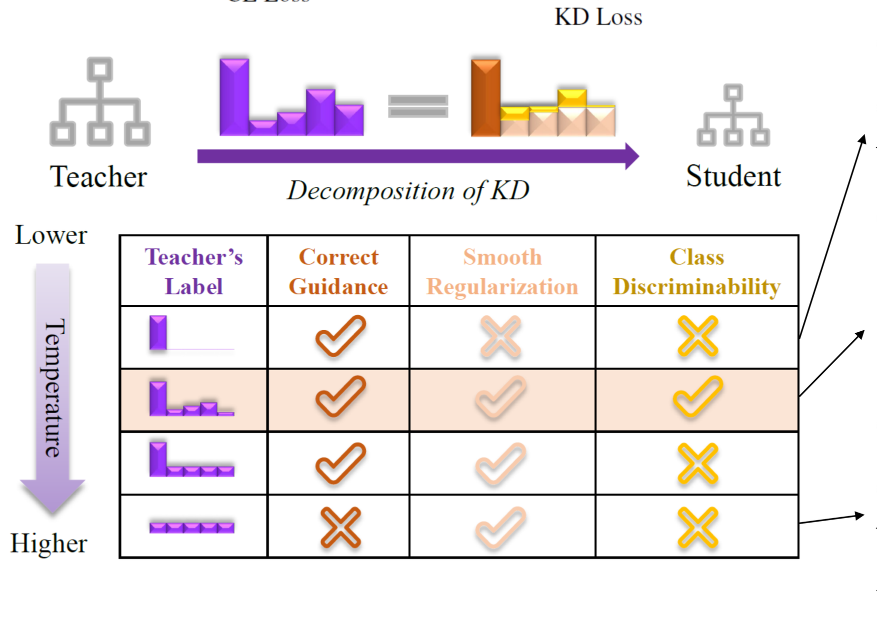
We focus on: **Why large networks may not teach well? How to make large networks teach better via simple methods?**




## Trial and Motivation

Why temperature $\tau$ of teacher net should be proper in traditional KD:

$$\ell = \underbrace{-(1-\lambda)\log \mathbf{p}_y^S(1)}_{\text{CE Loss}} \underbrace{-\lambda\tau^2 \sum_{c=1}^{C} \mathbf{p}_c^T(\tau)\log \mathbf{p}_c^S(\tau)}_{\text{KD Loss}}, \quad p^T(\tau) = SF(f;\tau)$$



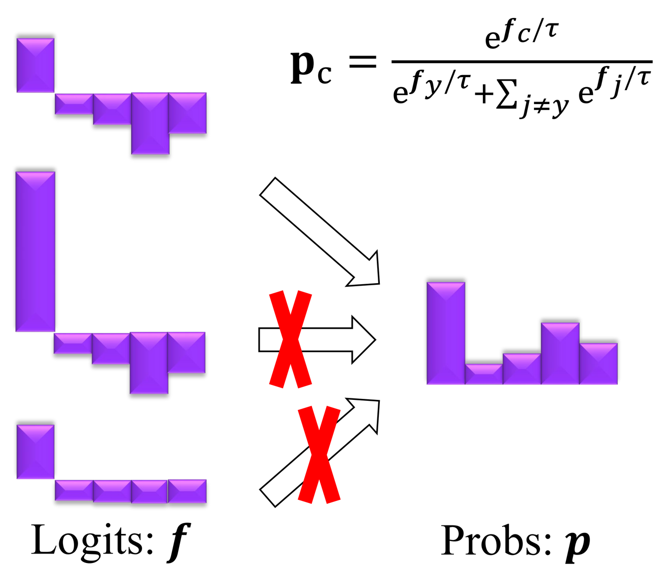

Too small temperature: Output of teacher nets tends to be One-Hot, which provides no extra information for student

**Proper temperature: Probs of wrong classes vary a lot**

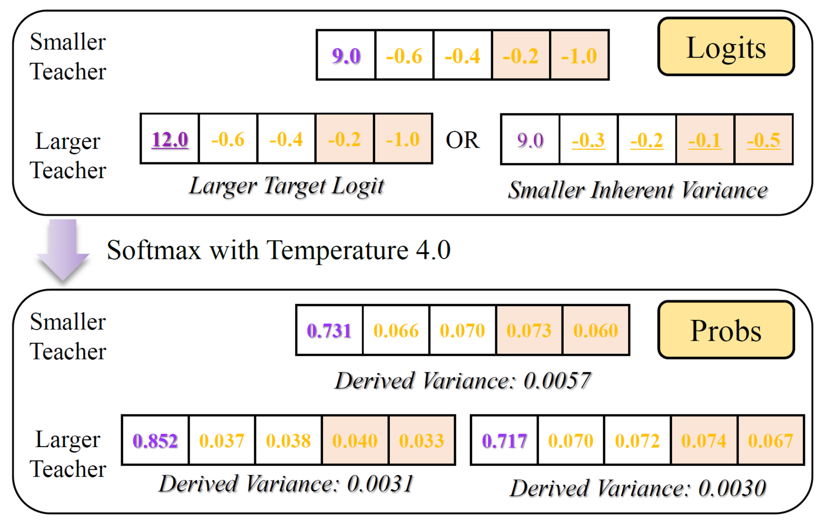
Too large temperature: Output of teacher nets tends to be uniform, which is similar to Label Smoothing

Guess: the reason why large nets cannot teach well lies in that probs of wrong classes cannot vary differently regardless of temperature

How to depict the distinctness of wrong classes: variance of probs of wrong classes




**Visual display**



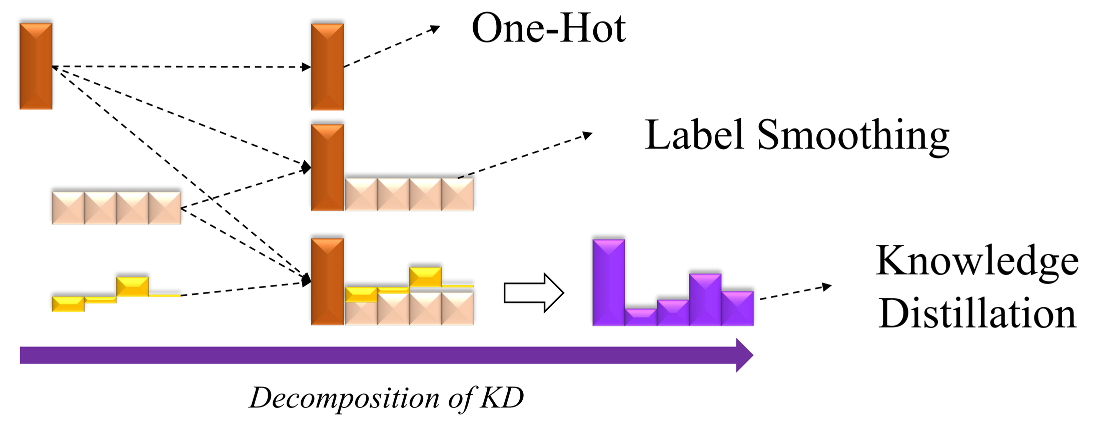

**Numerical example**

## Method

Decompose KD into three parts:

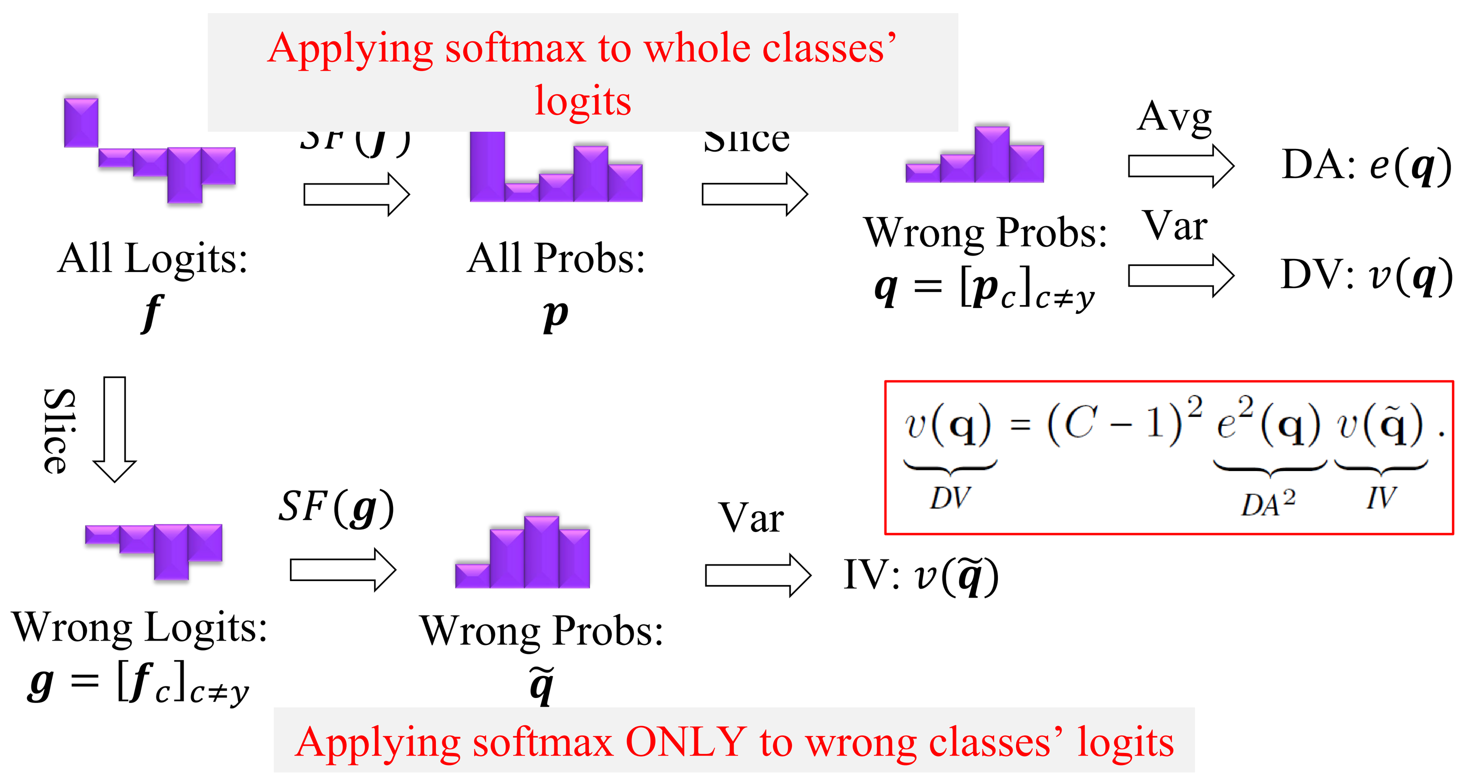
$$\ell = \underbrace{-(1-\lambda)\log \mathbf{p}_y^S(1)}_{\text{CE Loss}} \underbrace{-\lambda\tau^2 \sum_{c=1}^{C} \mathbf{p}_c^T(\tau)\log \mathbf{p}_c^S(\tau)}_{\text{KD Loss}},$$

$$\ell_{\text{kd}} = \underbrace{-\mathbf{p}_y^T(\tau)\log \mathbf{p}_y^S(\tau)}_{\text{Correct Guidance}} \underbrace{- \sum_{c\neq y} e\left(\mathbf{q}^T(\tau)\right)\log \mathbf{p}_c^S(\tau)}_{\text{Smooth Regularization}} \underbrace{- \sum_{c\neq y}\left(\mathbf{p}_c^T(\tau) - e\left(\mathbf{q}^T(\tau)\right)\right)\log \mathbf{p}_c^S(\tau)}_{\text{Class Discriminability}}.$$




**Proposition 4.4** (Derived Variance vs. Inherent Variance). *The derived variance is determined by the square of derived average and the inherent variance via:*

$$\underbrace{v(\mathbf{q})}_{DV} = (C-1)^2 \underbrace{e^2(\mathbf{q})}_{DA^2} \underbrace{v(\tilde{\mathbf{q}})}_{IV}. \quad (4)$$




Utilize this equation to explain why large net cannot teach well:

*Remark* 4.5. Fixing $\mathbf{g}$ and $\tau$, a higher target logit $\mathbf{f}_y$ leads to a higher $\mathbf{p}_y$, i.e., a smaller *derived average* $e(\mathbf{q})$.

*Remark* 4.6. Fixing $\tau$, less varied wrong logits $\mathbf{g}$ leads to less varied $\tilde{\mathbf{q}}$, i.e., a smaller *inherent variance* $v(\tilde{\mathbf{q}})$.

**Corollary 4.7.** *Suppose we have two teachers $T_1$ and $T_2$, and their logit vectors for a same sample are $\mathbf{f}^{T_1}$ and $\mathbf{f}^{T_2}$.*

- *If $\mathbf{f}_y^{T_1} \geq \mathbf{f}_y^{T_2}$ while $\mathbf{g}^{T_1}$ and $\mathbf{g}^{T_2}$ are nearly the same, then $\mathbf{p}_y^{T_1} \geq \mathbf{p}_y^{T_2}$ (Remark 4.5) while $v(\tilde{\mathbf{q}}^{T_1}) \approx v(\tilde{\mathbf{q}}^{T_2})$. Hence, $v(\mathbf{q}^{T_1}) \leq v(\mathbf{q}^{T_2})$.*
- *If $\mathbf{f}_y^{T_1} \approx \mathbf{f}_y^{T_2}$ while $v(\mathbf{g}^{T_1}) \leq v(\mathbf{g}^{T_2})$, then $\mathbf{p}_y^{T_1} \approx \mathbf{p}_y^{T_2}$ while $v(\tilde{\mathbf{q}}^{T_1}) \leq v(\tilde{\mathbf{q}}^{T_2})$ (Remark 4.6). Hence, $v(\mathbf{q}^{T_1}) \leq v(\mathbf{q}^{T_2})$.*

- Logit of correct class provided by large nets is quite large which leads to small DA.
- Logits of wrong classes provided by large net are less varied which leads to small IV.

Conclusion: **Large nets provide small DV. Traditional temperature scaling cannot make probs of wrong classes variant.**

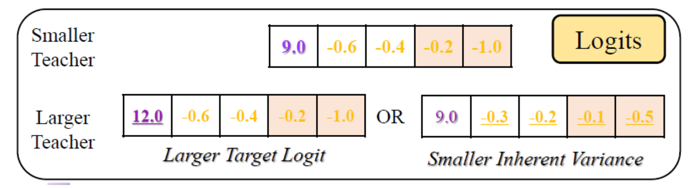
We propose Asymmetric Temperature Scaling (ATS):

$$\mathbf{p}_c(\tau_1,\tau_2) = \exp\left(\mathbf{f}_c/\tau_c\right) / \sum_{j\in[C]} \exp\left(\mathbf{f}_j/\tau_j\right), \quad \tau_i = \mathcal{I}\{i=y\}\tau_1 + \mathcal{I}\{i\neq y\}\tau_2, \forall i \in [C],$$

- If the teacher outputs a larger logit $\mathbf{f}_y$ for the correct class, a relatively larger $\tau_1$ could decrease it to a reasonable magnitude, i.e., decreasing $\mathbf{p}_y$ and increasing $e(\mathbf{q})$, and finally increasing the *derived variance* $v(\mathbf{q})$;
- If the teacher outputs less varied logits $\mathbf{g}$ for wrong classes, a relatively smaller temperature $\tau_2$ could make them more diverse, i.e., increasing $v(\tilde{\mathbf{q}})$, finally increasing the *derived variance* $v(\mathbf{q})$.

- Logit of correct class is large. Relatively larger $\tau_1$ could increase DA.
- Logits of wrong classes are similar. Relatively smaller $\tau_2$ could increase IV.

Conclusion: ATS can **enlarge DV provided by large nets to make probs of wrong classes more variant.**




Left ATS: $\tau_1 = 4.67, \tau_2 = 4.0$
Right ATS: $\tau_1 = 4.0, \tau_2 = 2.0$

## Experiments

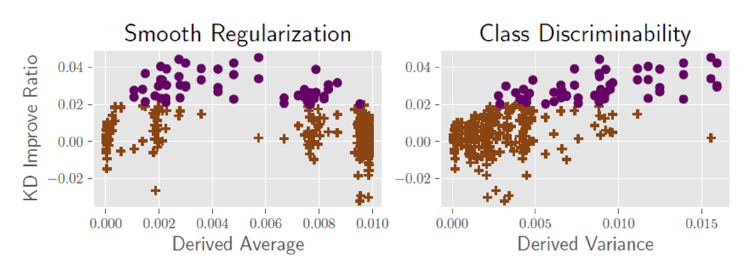



Figure 3: Correlations of *smooth regularization* (measured by *derived average*) and *class discriminability* (measured by *derived variance*) w.r.t. KD improvement ratio.

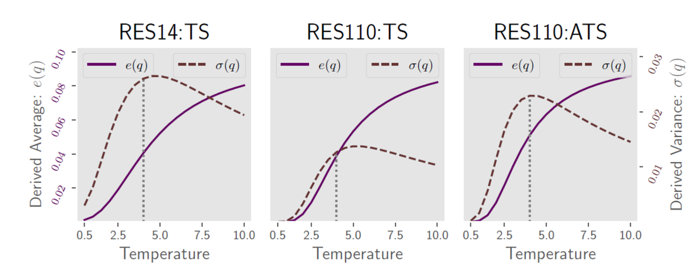
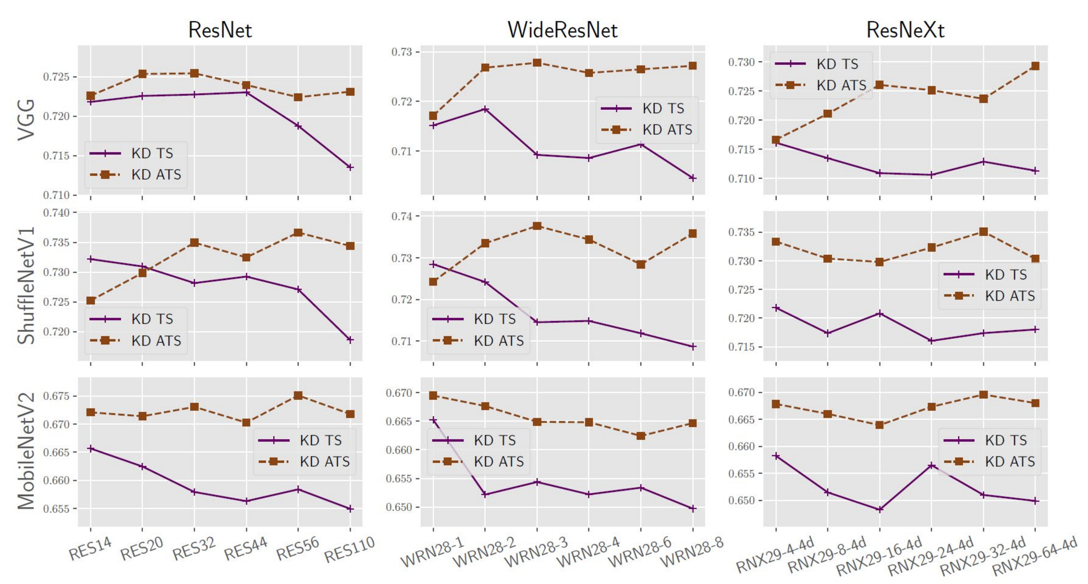



Figure 7: The change of *derived average* ($e(\mathbf{q})$) and *derived variance* ($v(\mathbf{q})$) as $\tau$ increases from 0.1 to 10.0 on CIFAR-10. The third one shows the results of ResNet110 with the proposed ATS. *DV* under TS is limited while ATS enlarges it.




In ATS, tune $\tau_1, \tau_2$ can make large nets teach better again.

Dataset: CIFAR-100
col: teacher net
row: student net
x-axis: capacity