Efficient Knowledge Distillation from Model Checkpoints

Chaofei Wang, Qisen Yang, Rui Huang, Shiji Song, Gao Huang

Department of Automation, Tsinghua University

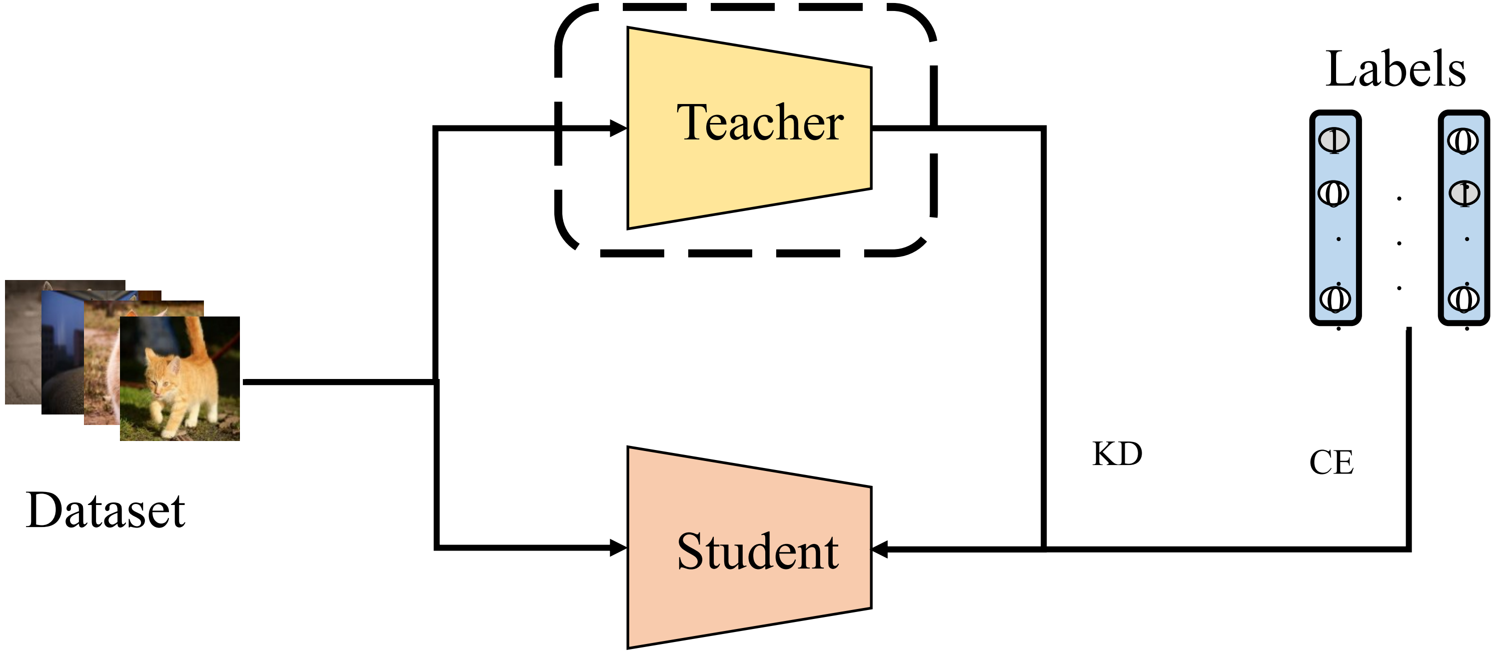
paper

清華大學
Tsinghua University

NEURAL INFORMATION PROCESSING SYSTEMS

Background

Knowledge distillation: train compact models (students) with the supervision of large and strong models (teachers).



Loss function: $L_{\text{KD}} = \alpha H(Y_{\text{true}}, P_{\text{S}}) + (1-\alpha) H(P^{\tau}_{\text{T}_{\text{full}}}, P^{\tau}_{\text{S}})$

CE KD

Background

Typical teachers: a well trained network or an ensemble of them.

$$L_{\mathrm{KD}} = \underbrace{\alpha H(Y_{\mathrm{true}}, P_{\mathrm{S}})}_{\text{CE}} + \underbrace{(1-\alpha) H(P_{\mathrm{T_{full}}}^{\tau}, P_{\mathrm{S}}^{\tau})}_{\text{KD}}$$

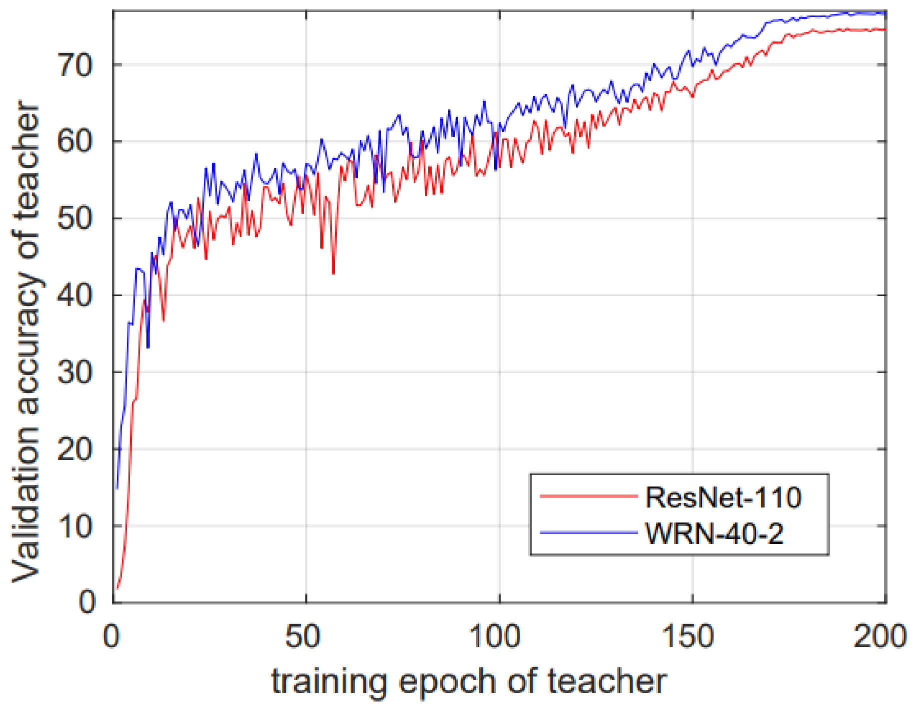
$$L_{\mathrm{EKD}} = \alpha H(Y_{\mathrm{true}}, P_{\mathrm{S}}) + (1-\alpha) H\left(\frac{1}{M}\sum_{i=1}^{M} P_{\mathrm{T_{full}}_i}^{\tau}, P_{\mathrm{S}}^{\tau}\right).$$

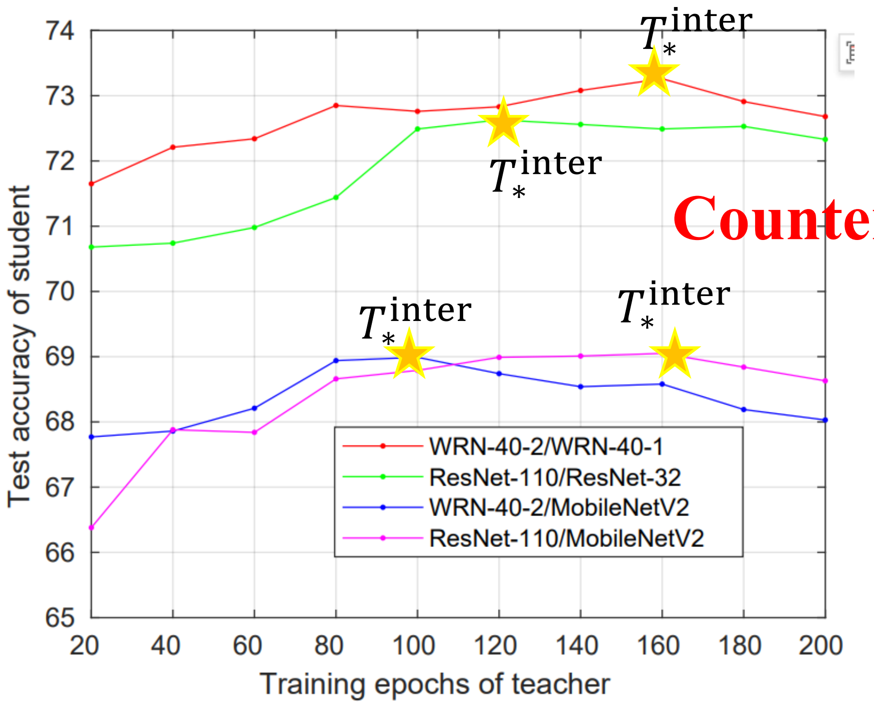
However, high performing models may not necessarily be good teachers.

An extreme example: if $P_{T^{\mathrm{full}}} \approx Y_{\mathrm{true}}$, KD would fail.

Exploratory Experiments

Intermediate Teacher vs. Full Teacher: The full teacher is a fully converged teacher model while the intermediate teacher is a checkpoint model in the training trajectory (e.g. half-trained model) .

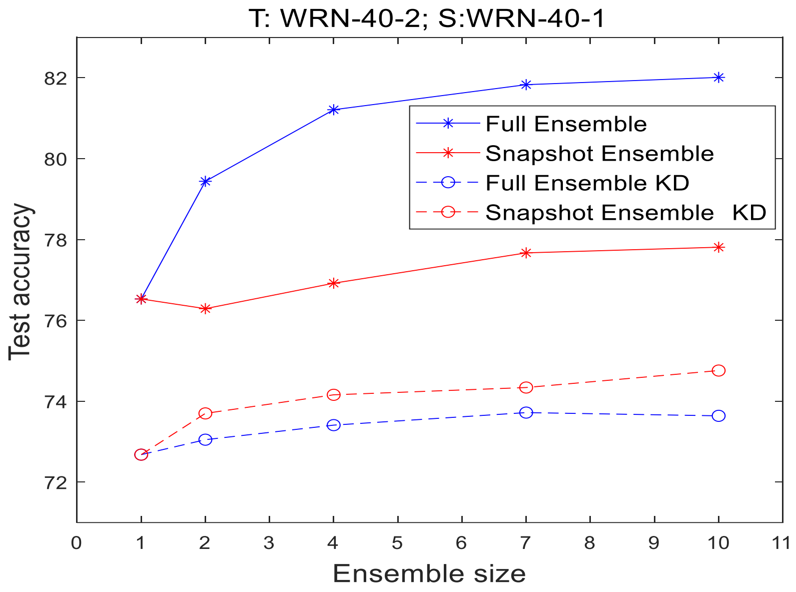




Counterintuitive!

Exploratory Experiments

Snapshot Ensemble[1] **vs. Full Ensemble**: The Full Ensemble is the standard ensemble of several independently trained full teacher models. The Snapshot Ensemble is an ensemble of several intermediate teacher models along the same optimization path.

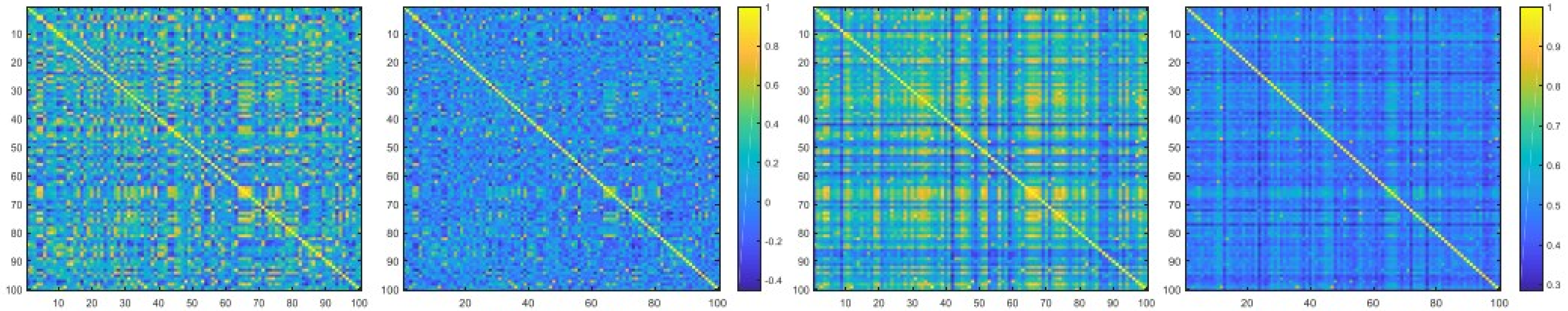


Counterintuitive!

[1] Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., & Weinberger, K. Q. (2017). Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*.

Why can intermediate models win?

Visualization: class correlation information of T^{inter} and T^{full}.



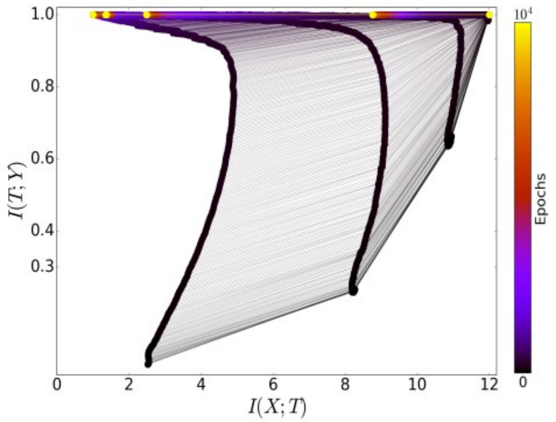
(a) Logits; Epoch = 100 (b) Logits; Epoch = 200 (c) Softmax τ = 5; Epoch = 100 (d) Softmax τ = 5; Epoch = 200

Observation: T^{inter} retains more class correlation information than T^{full}. For T^{full}, it is hard to reveal sufficient class correlation information by applying a high temperature to soften the network prediction.

Why can intermediate models win?

Information Bottleneck and Deep Neural Network

The optimization goal of DNN[2]: $\min_{F} \{I(X;F) - \beta I(F;Y)\}$



In the 1st stage: $I(X;F)\uparrow$
In the 2nd stage: $I(X;F)\downarrow$

Inference: a fully converged model tends to be overconfident and may already have collapsed representations for non-targeted classes.

[2] Shwartz-Ziv, R., & Tishby, N. (2017). Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810.

How to select the optimal model checkpoints?

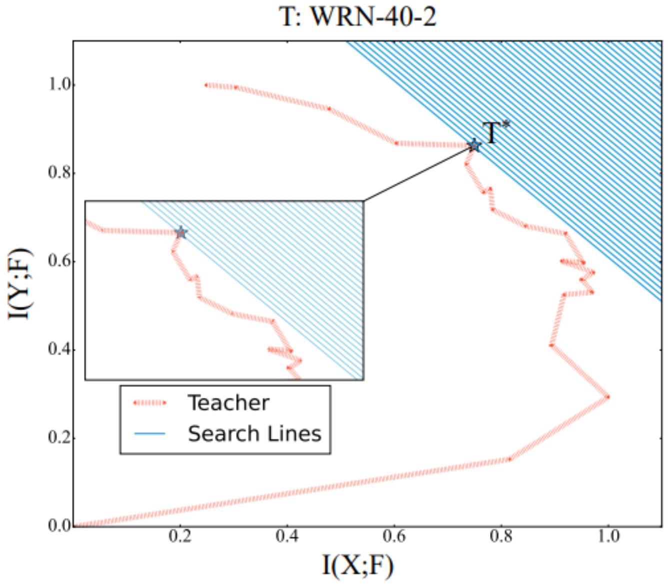
Solving the optimization problem:

$$\max_{F} \left\{ I(X;F) + I(Y;F) \right\},$$

where F belongs to the set of representations in intermediate teacher models.

Table 3: KD Results of the optimal intermediate models on CIFAR-100. The intermediate teacher models are selected at different epochs. The best results are **bold-faced**.

Network structure		Accuracy of T&S		KD accuracy of different intermediate teachers				
T	S	T	S	$T^{0.3}$	$T^{0.5}$	$T^{0.7}$	T^{full}	T^{*}
WRN-40-2	WRN-40-1	76.53	70.38	72.34±0.10	72.76±0.24	73.08±0.05	72.68±0.10	**73.26**±0.03
	MobileNetV2		64.49	68.21±0.33	**68.99**±0.12	68.54±0.07	68.03±0.34	68.58±0.34
ResNet-110	ResNet-32	73.41	70.16	70.74±0.18	72.49±0.32	72.46±0.30	72.48±0.22	**72.63**±0.13
	MobileNetV2		64.49	67.84±0.26	68.79±0.17	**69.01**±0.20	68.63±0.35	68.99±0.33
Average		74.97	67.38	69.78	70.76	70.77	70.46	**70.87**



Take-aways

- Enriching the “dark knowledge” of the teacher is more important than Improving the performance of the teacher.
- $T^{0.5}$ is generally can be an more efficient teacher than T^{full}.
- Snapshot Ensemble can be an more efficient teacher than Full Ensemble.
- $I(X;F_t)$ can be used to explain the “dark knowledge”. More $I(X;F_t)$ is the key reason that T^{inter} can beat T^{full}.

.

Thanks！