SPARSE LOGITS SUFFICE TO FAIL KNOWLEDGE DISTILLATION

PAIR²STRUCT@ICLR2022

Haoyu Ma[1], Yifan Huang[2], Hao Tang[1], Chenyu You[3], Deying Kong[1], Xiaohui Xie[1]
[1]University of California, Irvine, [2]Southeast University, [3]Yale University

Code is available

Motivations

❖ **To solve the problem of intellectual properties (IPs) leakage caused by knowledge distillation, *Nasty Teacher*[1] retrains a new model and distorting its output distribution from the normal one via an adversarial loss. However, it is unclear why the distorted distribution is catastrophic to the student model, as the nasty logits still maintain the correct labels.**

Methodology

❖ **Sparse probability: When the student learns from the sparse probabilities $p_\tau^{\sim T}(k)$, the KL divergence in knowledge distillation and loss funcation are rewritten as:**

$$\begin{aligned}\mathcal{KL}(\tilde{p}_\tau^T, p_\tau^S) &= -\sum_{k=1}^{K} \tilde{p}_\tau^T(k)\log p_\tau^S(k) = -\sum_{k\in\mathbf{M}} (p_\tau^T(k)+\delta(k))\log p_\tau^S(k) \\ &\approx -\sum_{k\in\mathbf{M}} p_\tau^T(k)\log p_\tau^S(k) - \frac{1-r}{r}\sum_{k\in\mathbf{M}}\frac{1}{K}\log p_\tau^S(k) \\ &\approx -\frac{1}{rK}\sum_{k\in\mathbf{M}}\log p_\tau^S(k)\end{aligned}$$

$$\tilde{\mathcal{L}}_{KD} = (1-\alpha)\mathcal{H}(p^S, y) + \frac{\alpha\tau^2}{rK}\left[-\sum_{k\in\mathbf{M}}\log p_\tau^S(k)\right]$$

Compared with learning from the hard label, the studentmodel cannot identify the difference between categories within the subset *M* and undoubtedly give a wrong prediction.

❖ **Stingy teacher: We propose a new method that directly manipulates the output logits of any pre-trained model to achieve the effect of the nasty teacher, named *Stingy teacher*. The logit still maintains the similarity structure among categories, but it is "stingy" as it only provides the information of a few categories.**

❖ **Implementation: Given the logits z_k^T from the pre-trained model, the stingy logit z_i^{ST} still keep the value z_k^T if k is in the top-N subset M^{ST}. Otherwise, it is set to negative infinity.**

References

[1] Haoyu Ma, Tianlong Chen, Ting-Kuei Hu, Chenyu You, Xiaohui Xie, and Zhangyang Wang. Undistillable: Making a nasty teacher that cannot teach students. In ICLR, 2021b.

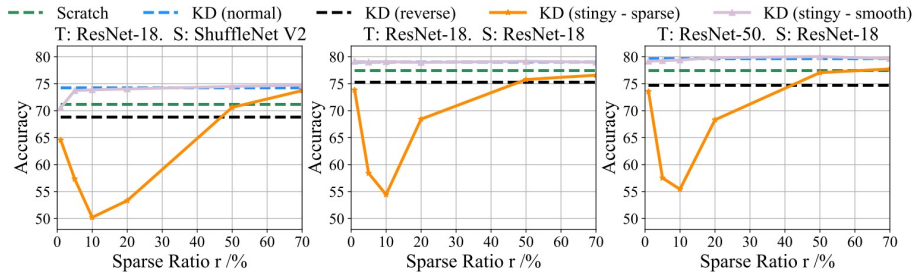
Results

❖ **Standard KD**

Table 1: Comparison of the nasty teacher and the stingy teacher on CIFAR-100.

Teacher network	Teacher accuracy	Students accuracy after KD			
		Shufflenetv2	MobilenetV2	ResNet-18	Teacher Self
Student baseline	-	71.17	69.12	77.44	-
ResNet-18 (normal)	77.44	74.24 (+3.07)	73.11 (+3.99)	79.03 (+1.59)	79.03 (+1.59)
ResNet-18 (nasty)	77.42 (-0.02)	64.49 (-6.68)	3.45 (**-65.67**)	74.81 (-2.63)	74.81 (-2.63)
ResNet-18 (stingy)	77.44 (**-0.00**)	50.22 (**-20.95**)	6.78 (-62.34)	54.44 (**-23.00**)	54.44 (**-23.00**)
ResNet-50 (normal)	78.12	74.00 (+2.83)	72.81 (+3.69)	79.65 (+2.21)	80.02 (+1.96)
ResNet-50 (nasty)	77.14 (-0.98)	63.16 (-8.01)	3.36 (**-65.76**)	71.94 (-5.50)	75.03 (-3.09)
ResNet-50 (stingy)	78.12 (**-0.00**)	49.05 (**-22.12**)	5.52 (-63.60)	55.44 (**-22.00**)	55.63 (**-22.49**)
ResNeXt-29 (normal)	81.85	74.50 (+3.33)	72.43 (+3.31)	80.84 (+3.40)	83.53 (+1.68)
ResNeXt-29 (nasty)	80.26 (-1.59)	58.99 (-12.18)	1.55 (**-67.57**)	68.52 (-8.92)	75.08 (-6.77)
ResNeXt-29 (stingy)	81.85 (**-0.00**)	49.46 (**-21.71**)	6.93 (-62.19)	58.70 (**-18.74**)	54.18 (**-27.67**)

❖ **Comparison of KD from three types of logits: the "stingy-sparse", the "stingy-smooth", and the "reversed logits". Experiments are conducted on CIFAR-100.**



❖ **Data-Free KD**

Table 2: Data-free KD from nasty teacher on CIFAR-10 and CIFAR-100

dataset	CIFAR-10		CIFAR-100	
Teacher Network	Teacher Accuracy	DAFL	Teacher Accuracy	DAFL
ResNet34 (normal)	95.42	92.49	76.97	71.06
ResNet34 (nasty)	94.54 (-0.88)	86.15 (-6.34)	76.12 (-0.79)	65.67 (-5.39)
ResNet34 (stingy)	95.42 (-0.00)	90.26 (-2.23)	76.97 (-0.00)	69.02 (-2.04)

Contact: {haoyum3, htang6, deyingk, xhx}@uci.edu, yifanhuang@seu.edu.cn, chenyu.you@yale.edu