SPARSE LOGITS SUFFICE TO FAIL KNOWLEDGE DISTILLATION

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

ABSTRACT

Knowledge distillation (KD) aims to transfer the power of pre-trained teacher models to (more lightweight) student models. However, KD also poses the risk of intellectual properties (IPs) leakage of teacher models. Even if the teacher model is released as a black box, it can still be cloned through KD by imitating input-output behaviors. To address this unwanted effect of KD, the concept of Nasty Teacher was proposed recently. It is a special network that achieves nearly the same accuracy as a normal one, but significantly degrades the accuracy of student models trying to imitate it. Previous work builds the nasty teacher by retraining a new model and distorting its output distribution from the normal one via an adversarial loss. With this design, the "nasty" teacher tends to produce sparse and noisy logits. However, it is unclear why the distorted distribution is catastrophic to the student model, as the nasty logits still maintain the correct labels. In this paper, we provide a theoretical analysis of why the sparsity of logits is key to Nasty Teacher. Furthermore, we propose an ideal version of nasty teacher to prevent imitation through KD, named *Stingy Teacher*. The Stingy Teacher directly manipulates the logits of a standard pre-trained network by maintaining the values for a small subset of classes while zeroing out the rest. Extensive experiments on several datasets demonstrate that stingy teacher is more catastrophic to student models on both standard KD and data-free KD. Source code and trained model can be found at https://https://github.com/HowieMa/stingy-teacher.

1 INTRODUCTION

Knowledge Distillation (KD) (Hinton et al., 2015) aims to transfer the ability of a pre-trained network (teacher) to another network (student). It has been widely applied in many areas including image classification Hinton et al. (2015); Ma et al. (2021a); Chen et al. (2021a;b), object detection Wang et al. (2019); Zheng et al. (2021), semantic segmentation Liu et al. (2019); You et al. (2022) and speech recognition Oord et al. (2018); You et al. (2021a;b;c). Typically, the teacher model is more sophisticated with higher performance. The performance of lightweight student model is boosted by imitating the output logits (Hinton et al., 2015; Park et al., 2019; Mirzadeh et al., 2020; Furlanello et al., 2018; Zhang et al., 2019; Yuan et al., 2020) or intermediate activation maps (Romero et al., 2014; Zagoruyko & Komodakis, 2016; Passalis & Tefas, 2018; Ahn et al., 2019; Li et al., 2020) from teacher models.

Recent work (Ma et al., 2021b) suggests that KD, on the other hand, poses the risk of exposing intellectual properties (IP). Even if the trained model is released as "black boxes", it can still be cloned by imitating the input-output behaviors, as some data-free KD methods (Lopes et al., 2017; Chen et al., 2019; Nayak et al., 2019; Yin et al., 2020; Truong et al., 2021; Yin et al., 2021) eliminate the necessity of having access to the original training examples. To alleviate this side effect of KD, (Ma et al., 2021b) introduces the concept of the *Nasty Teacher*: a specially trained teacher network that yields nearly the same performance as a normal one, but significantly degrades the performance of student models trying to imitate it. To this end, (Ma et al., 2021b) proposes to obtain special logits via adversarial training that maximizes the difference between the output of the nasty teacher and a normal pre-trained (teacher) network. With this design, the accuracy of the student learned from the nasty teacher can be degraded by over 10%.

However, several issues remain unsolved in the aforementioned Nasty Teacher approach. First, it is unclear why changing the distribution of the output probabilities is catastrophic to the student model. Although the logits are noisy, it still has the correct output label. The student should not be significantly degraded as long as it learns the teacher well. Second, the accuracy of the Nasty Teacher model can also drops over 2%, which is unacceptable in many applications. In this case, the protection of IP comes at the cost of the accuracy of the model. One may need to carefully balance the pros and cons of the nasty teacher, which undermines its utility.

In this paper, we empirically validate that the sparsity of logits is key to the nasty teacher, for which we also provide a theoretical analysis to understand when sparse logits can be useful to degrade the performance of student networks. Contrary to the common belief, we find out that the logits do not have to be very noisy (in which case the teacher becomes "ignorant" itself) – as long as the teacher supplies sparse logits, the student model will suffer. Based on this empirical observation, we propose to construct the *Stingy Teacher*, an *ideal* version of the nasty teacher to prevent knowledge leaking and unauthorized model cloning. The stingy teacher directly manipulates the logits of a pre-trained network by keeping the values for the classes with relatively high probabilities, and zeroing out the rest. Such special sparse logits can still preserve the teacher's accuracy (hence the teacher itself is still "knowledgeable"), as well as the partial inter-class similarity structure. However, it is *"stingy"* and refuses to provide full information of all classes. This simple design is innocuous to the original trained model and requires no retraining, as we just manually re-shape its logits without touching pre-trained weights. Thus, it is easy to be applied to any huge networks in real applications. Although ideally, we believe the property of the stingy teacher is potentially helpful for designing loss function in the future work. We summarize our contributions as follows:

- We provide a theoretical understanding of why introducing sparsity to the output probabilities makes KD ineffective to distill knowledge from the teacher model.
- We propose a simpler yet more effective Nasty Teacher called *Stingy Teacher*. It directly manipulates the logits by keeping the values for the top-N classes and zeroing out the rest. Thus, there is no accuracy drop for the teacher models.
- Extensive experiments on both standard KD and data-free KD demonstrate that the Stingy Teacher can make the student model learned from it fails substantially in terms of accuracy.

2 PRELIMINARIES

Knowledge Distillation The key idea of KD (Hinton et al., 2015) is to force the student network to imitate the input-output behavior of pre-trained teacher networks. Suppose there are K classes in total, the student model S produces the soft probability of each label $k \in \mathbf{C} = \{1, 2, ..., K\}$: $p_\tau^S(k) = \sigma_\tau(z_k^S)$, where z_k^S is the logit from S and $\sigma_\tau(\cdot)$ is the scaled Softmax function with temperature τ. Similarly, denote $p_\tau^T(k) = \sigma_\tau(z_k^T)$ as the soft output from the pre-trained teacher network T. The student S is trained by minimizing the cross-entropy loss $\mathcal{H}(\cdot, \cdot)$ and the K-L divergence $\mathcal{KL}(\cdot, \cdot)$ between the student and teacher predictions:

$$\mathcal{L}_{KD} = \alpha\tau^2\mathcal{KL}(p_\tau^T, p_\tau^S) + (1-\alpha)\mathcal{H}(p_{\tau=1}^S(k), y). \tag{1}$$

Nasty Teacher The *Nasty Teacher* (Ma et al., 2021b) is a defensive approach for model owners to alleviate the issue of model cloning through KD. By definition, the accuracy of the nasty teacher is the same as its normal one, while any arbitrary student networks who attempt to imitate it will be degraded. The nasty teacher NT propose a strategy to maintain the correct class assignments, while disturbs its in-correct class assignments. In detail, it is trained from scratch by simultaneously minimizing the cross-entropy loss with the hard label and maximizing the K-L divergence with the pre-trained normal teacher network T:

$$\mathcal{L}_{NT} = \mathcal{H}(p^{NT}, y) - \omega\tau^2\mathcal{KL}(p_\tau^T, p_\tau^{NT}), \tag{2}$$

where ω is the weight to control the trade-off between performance suffering and nasty behavior.

3 METHODOLOGY

3.1 SPARSE PROBABILITIES: KEY TO THE SUCCESS OF NASTY TEACHER

The previous work (Ma et al., 2021b) hypothesizes that the noisy responses of NT give a false sense of generalization, and thus degrade the accuracy of student models. However, even if the

"dark knowledge" encoded in the output is disturbed, the output still maintains the (almost) correct predictions. The student network should give a reasonable prediction as long as it perfectly mimics the disturbed logits. Thus, we question whether the noise is the major effect which results in the accuracy drop of the student.

Besides noise, we find that the probability distribution produced by the nasty teacher also yields another interesting property, the *Sparsity*. When increasing the probability of some incorrect categories, the nasty logits meanwhile reduce or even zero out the probabilities of the rest categories. Thus, the nasty logits are more likely to be sparse labels, rather than a smooth distribution. Our question of curiosity is hence: "*Is sparsity the key to the nasty teacher*"?

In this section, we provide a mathematical analysis to understand why the student model will be degraded when imitating the sparse probabilities, whether it is noisy or not. Denote the sparse probabilities as $\tilde{p}_\tau^T(k)$. Compared with the original distribution $p_\tau^T(k)$, we only preserve the probabilities of a subset $\mathbf{M}$ ($\mathbf{M} \subset \mathbf{C}$) of categories, while setting the probabilities of the rest categories to 0. The label of the top-1 prediction is always preserved in $\mathbf{M}$. Specifically, we define the adjusted probability as $\tilde{p}_\tau^T(k) = p_\tau^T(k) + \delta(k)$ if $k \in \mathbf{M}$, and 0 otherwise. The $\delta(k)$ is added to ensure that the adjusted probabilities are properly normalized (i.e., $\sum_k \tilde{p}_\tau^T(k) = 1$). Let $N = |\mathbf{M}|$ with $1 \leq N < K$. We define $r = \frac{N}{K}$ as the sparse ratio of $\tilde{p}_\tau^T(k)$. Moreover, (Ma et al., 2021b) finds that the accuracy of the student network is much worse when imitating the nasty teacher with a larger τ. Typically, the output will be very soft and similar to a uniform distribution with a large τ (Hinton et al., 2015), especially when the total number of class K is large. Therefore, we assume that all $p_\tau^T(k)$ (except for the top-1 prediction) are equal when τ is relatively large, and apply a uniform distribution to approximate them for simplicity. Let j be the class of the top-1 prediction, $p_\tau^T(k)$ can be approximated by:

$$p_\tau^T(k) \approx \begin{cases} \frac{1}{K} - \epsilon, & \text{if } k \neq j \\ \frac{1}{K} + \epsilon(K-1), & \text{if } k = j \end{cases} \tag{3}$$

where ϵ is sufficiently small ($0 < \epsilon \ll \frac{1}{K}$). Thus, the residual value $\delta(k)$ can be approximated as $\delta(k) \approx \frac{1-r}{rK}$ for all $k \in \mathbf{M}$. The detailed derivation process is presented in Appendix A1. When the student learns from the sparse probabilities $\tilde{p}_\tau^T(k)$, the KL divergence in Eq. 1 is rewritten as [1] :

$$\begin{aligned} \mathcal{KL}(\tilde{p}_\tau^T, p_\tau^S) &= -\sum_{k=1}^{K} \tilde{p}_\tau^T(k) \log p_\tau^S(k) = -\sum_{k \in \mathbf{M}} (p_\tau^T(k) + \delta(k)) \log p_\tau^S(k) \\ &\approx -\sum_{k \in \mathbf{M}} p_\tau^T(k) \log p_\tau^S(k) - \frac{1-r}{r} \sum_{k \in \mathbf{M}} \frac{1}{K} \log p_\tau^S(k) \\ &\approx -\frac{1}{rK} \sum_{k \in \mathbf{M}} \log p_\tau^S(k) \end{aligned} \tag{4}$$

For the first approximation, we replace $\delta(k)$ with $\frac{1-r}{rK}$, and for the second approximation, we replace $p_\tau^T(k)$ with $\frac{1}{K}$. Thus, when learning from the sparse logits, the loss function in Eq. 1 is rewritten as:

$$\tilde{\mathcal{L}}_{KD} = (1-\alpha)\mathcal{H}(p^S, y) + \frac{\alpha\tau^2}{rK} \left[-\sum_{k \in \mathbf{M}} \log p_\tau^S(k) \right] \tag{5}$$

Compared with learning from the hard label, the second term in Eq. 5 equally maximizes the probabilities of all classes within the subset $\mathbf{M}$. This term forces the model to produce high responses on all categories within the subset $\mathbf{M}$. A sparse logit (i.e., $r < 0.1$) leads to a large weight $1/r$, making the student model spend more effort to optimize the second term. Consequently, the student model cannot identify the difference between categories within the subset $\mathbf{M}$ and undoubtedly give a wrong prediction. Similarly, a larger α or τ leads to the same effect.

3.2 STINGY TEACHER

Based on the above theoretical analysis, 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 stingy teacher only keeps the information for the top-N classes, while zeroing out the rest. Thus,

[1] we intend to discard the entropy of $\tilde{p}_\tau^T(k)$ as it does not contribute to the gradient of student S

the logit still maintains the similarity structure among categories, but it is "stingy" as it only provides the information of a few categories. 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 $\mathbf{M}^{ST}$. Otherwise, it is set to negative infinity.

4 EXPERIMENTS

4.1 STINGY TEACHER ON STANDARD KNOWLEDGE DISTILLATION

In the standard KD, the student has the access to the original training examples. We follow all experimental settings in (Ma et al., 2021b) to explore our stingy teacher on standard KD.

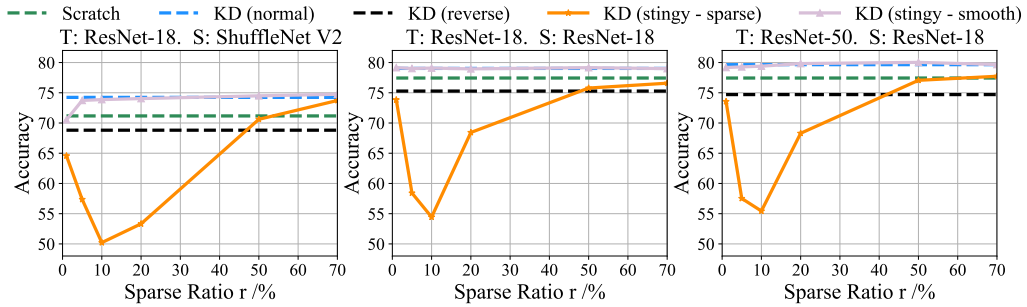

Figure 1: Comparison of KD from three types of logits: the "stingy-sparse", the stingy-smooth, and the "reversed logits". Experiments are conducted on CIFAR-100.

4.1.1 THE EFFECT OF SPARSE LOGITS

We firstly compare the stingy logits (*"stingy - sparse"*) with two variants. 1) "*stingy - smooth*": We keep the same subset of logits, but replace the rest of logits with their average; 2) "*reversed logits*": We keep the top-1 prediction, but reverse the value of the rest of the logits. Both of them still maintain the smoothness property. We explore the relationship between the sparse ratio r and the accuracy of student networks distilling from each types of logits. Results are presented in Fig. 1. Firstly, when the logits is smooth, even if the dark knowledge is limited, the student can still obtain some improvements. This also supports that KD plays the role of label smooth regularization (Yuan et al., 2020). Secondly, when the logits are misleading, the accuracy of student can be downgraded 5% to 8%. When the capacity of student is huge, the damage is mitigated. This suggests that a noisy logits is somewhat harmful for lightweight student networks. Thirdly, when the logit is sparse, the accuracy of the student model is significantly degraded, whatever the capacity of the student model. When sparse ratio r is around 10%, most students achieve the worst performance, i.e, more than 20% accuracy drop. All of the experiments support our claim that sparsity is the major reason leads to the accuracy drop of student networks, compared with noise.

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**)

4.1.2 COMPARISON WITH NASTY TEACHER

We apply the surprising property of sparse logits to the standard KD and compare it with the nasty teacher. We empirically set the sparse ratio to 0.1. Table 1 show the results on CIFAR-100. Results on CIFAR-10, Tiny-ImageNet and ImageNet are presented in Appendix A2. The performance of the stingy teacher always matches that of the normal teacher perfectly, as we only manipulate the logits. Meanwhile, the accuracy of students can be further degraded when distilling from the stingy teacher. Moreover, stingy teacher is more catastrophic to large student networks. In conclusion, the stingy teacher can significantly downgrade the performance of any network trying to clone it without sacrificing its own accuracy.

4.2 STINGY TEACHER ON DATA-FREE KNOWLEDGE DISTILLATION

In practice, as long as the student cannot perform well via the standard KD, the attacker can just train the model solely with the hard label. Instead, KD without accessing any training sample is a more realistic way, where the student can only access the output distributions from the teacher. In this way, the teacher is released as "black boxes". Following Ma et al. (2021b), we explore the stingy teacher on one popular data-free KD method, i.e., DAFL Chen et al. (2019). Results are shown in Table 2. Noticeably, the accuracy of students can still be downgraded up to 2% when distilling from the stingy teacher. Thus, the sparse logits also have the ability to downgrade the data-free KD. Although the nasty teacher can further destroy the student, it achieves this at the cost of accuracy drop, while the stingy teacher can exactly maintain the origin accuracy.

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)

4.3 DISCUSSIONS

We acknowledge that the stingy teacher is a relatively ideal version of the nasty teacher, as the logits are obtained in a post-processing way, rather than obtained from the vanilla output. On standard KD, we aim to use the stingy logits to reveal the effect of sparsity on the success of the nasty teacher. Experiment results suggest that sparsity is more effective than noise. We believe this interesting property of nasty logits can help the community to design extra loss functions or training strategies to make the model directly produce better nasty logits. On data-free KD, since the teacher is already released as a "black box", applying a post-processing method on the origin output is a reasonable way. It is true that some naive post-processing strategies such as releasing the top-1 / top-N categories can avoid distilling. Nevertheless, many practical cloud APIs, such as Google Vision AI, prefer to provide the probabilities to customers. In this case, the stingy logits can maximally maintain the property of the origin logits, such as the relative relationships among the top categories. To this end, the stingy teacher does not hurt the normal usage of the black-box API, while still has the ability to avoid distillation.

5 CONCLUSION

In this paper, we demonstrate that the sparsity of the logits is the main reason for the accuracy drop of student networks in the setting of the nasty teachers. Based on this property, we propose the stingy teacher. The stingy teacher is simple yet effective, which is implemented by keeping a small subset of top logits and zeroing out the rest. Extensive experiments demonstrate that our method is more catastrophic to student networks on standard KD and also effective on data-free KD. We believe this relatively ideal version of the nasty teacher can motivate the community to design better model protection strategies in the future.

REFERENCES

Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *CVPR*, pp. 9163–9171, 2019.

Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *ICCV*, pp. 3514–3522, 2019.

Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Long live the lottery: The existence of winning tickets in lifelong learning. In *ICLR*, 2021a.

Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *ICLR*, 2021b.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, pp. 1607–1616, 2018.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Guilin Li, Junlei Zhang, Yunhe Wang, Chuanjian Liu, Matthias Tan, Yunfeng Lin, Wei Zhang, Jiashi Feng, and Tong Zhang. Residual distillation: Towards portable deep neural networks without shortcuts. *NeurIPS*, 2020.

Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *CVPR*, pp. 2604–2613, 2019.

Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535*, 2017.

Haoyu Ma, Tianlong Chen, Ting-Kuei Hu, Chenyu You, Xiaohui Xie, and Zhangyang Wang. Good students play big lottery better. *arXiv preprint arXiv:2101.03255*, 3, 2021a.

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.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI*, pp. 5191–5198, 2020.

Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Zero-shot knowledge distillation in deep networks. In *ICML*, pp. 4743–4751, 2019.

Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. Parallel wavenet: Fast high-fidelity speech synthesis. In *ICML*, pp. 3918–3926, 2018.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, pp. 3967–3976, 2019.

Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *ECCV*, pp. 268–284, 2018.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Jean-Baptiste Truong, Pratyush Maini, Robert J Walls, and Nicolas Papernot. Data-free model extraction. In *CVPR*, pp. 4771–4780, 2021.

Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *CVPR*, pp. 4933–4942, 2019.

Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *CVPR*, pp. 8715–8724, 2020.

Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. See through gradients: Image batch recovery via gradinversion. In *CVPR*, pp. 16337–16346, 2021.

Chenyu You, Nuo Chen, and Yuexian Zou. Contextualized attention-based knowledge transfer for spoken conversational question answering. In *INTERSPEECH*, 2021a.

Chenyu You, Nuo Chen, and Yuexian Zou. Knowledge distillation for improved accuracy in spoken question answering. In *ICASSP*, 2021b.

Chenyu You, Nuo Chen, and Yuexian Zou. MRD-Net: Multi-Modal Residual Knowledge Distillation for Spoken Question Answering. In *IJCAI*, 2021c.

Chenyu You, Yuan Zhou, Ruihan Zhao, Lawrence Staib, and James S Duncan. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 2022.

Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *CVPR*, pp. 3903–3911, 2020.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *ICCV*, pp. 3713–3722, 2019.

Wu Zheng, Weiliang Tang, Li Jiang, and Chi-Wing Fu. Se-ssd: Self-ensembling single-stage object detector from point cloud. In *CVPR*, pp. 14494–14503, 2021.

A1 DETAILED MATHEMATICAL DERIVATION

In this appendix, we give detailed derivation process of Eq. 4. When the temperature τ is high, we can use a uniform distribution to approximate $p_\tau^T(k)$. As in Eq. 3, the soft probabilities $p_\tau^T(k)$ can be approximated with:

$$p_\tau^T(k) \approx \begin{cases} \frac{1}{K} - \epsilon, & \text{if } k \neq j \\ \frac{1}{K} + \epsilon(K-1), & \text{if } k = j \end{cases} \tag{A1}$$

Where ϵ is a neglectable factor ($0 \leq \epsilon \ll \frac{1}{K}$), and j is the original top-1 prediction. The new sparse logits $\tilde{p}_\tau^T(k)$ is defined with:

$$\tilde{p}_\tau^T(k) \approx \begin{cases} p_\tau^T(k) + \delta(k), & \text{if } k \in \mathbf{M} \\ 0, & \text{if } k \notin \mathbf{M} \end{cases} \tag{A2}$$

Once zeroing out the probabilities of class $k \notin \mathbf{M}$, the total discarded probabilities $\mathcal{P}_1$ can be derived by:

$$\mathcal{P}_1 = (K-N)p_\tau^T(k) = (K-N)\left(\frac{1}{K} - \epsilon\right) = (K - rK)\left(\frac{1}{K} - \epsilon\right) = (1-\epsilon K)(1-r) \tag{A3}$$

Again, $r = \frac{N}{K}$ is the sparse ratio, and N is the total number of element in the subset $\mathbf{M}$. Thus, the total preserved probabilities of class $k \in \mathbf{M}$ is

$$\mathcal{P}_2 = 1 - \mathcal{P}_1 = 1 - (1-\epsilon K)(1-r) = r + \epsilon K(1-r) \tag{A4}$$

To ensure the sum of $\tilde{p}_\tau^T(k)$ is equal to 1, we need re-normalize the probabilities of the preserved categories $k \in \mathbf{M}$. Here we just consider the simplest additional way, and distribute $\mathcal{P}_1$ onto class $k \in \mathbf{M}$ based on the original probability $p_\tau^T(k)$. Thus, the additional term $\delta(k)$ can be writen as

$$\delta(k) = \frac{p_\tau^T(k)}{\mathcal{P}_2}\mathcal{P}_1 \tag{A5}$$

Specifically, when $k \neq j$, we have

$$\delta(k) = \frac{p_\tau^T(k)}{\mathcal{P}_2}\mathcal{P}_1 = \frac{\frac{1}{K} - \epsilon}{r + \epsilon K(1-r)}(1-\epsilon K)(1-r) = \frac{(1-\epsilon K)^2}{r + \epsilon K(1-r)}\frac{1-r}{K} \tag{A6}$$

Since $\epsilon \ll \frac{1}{K}$, we have $0 \leq \epsilon K \ll 1$, thus $(1-\epsilon K)^2 \approx 1$. As $r < 1$, $r + \epsilon K(1-r) < r + \epsilon K \approx r$. Therefore, we can approximate $\delta(k)$ with

$$\delta(k) = \frac{(1-\epsilon K)^2}{r + \epsilon K(1-r)}\frac{1-r}{K} \approx \frac{1-r}{rK} \tag{A7}$$

When $k = j$, we have

$$\delta(j) = \mathcal{P}_1 - (N-1)\delta(k)_{k\in\mathbf{M}, k\neq j} \approx (1-r) - (rK-1)\frac{1-r}{rK} = \frac{1-r}{rK} \tag{A8}$$

In conclusion, we can approximate $\delta(k)$ with $\dfrac{1-r}{rK}$ for any $k \in \mathbf{M}$.

A2 MORE EXPERIMENTAL RESULTS

A2.1 COMPARISON WITH NASTY TEACHER

Besides CIFAR-100, we also compare the stingy teacher with the nasty teacher on CIFAR-10 and Tiny-ImageNet. Results are presented in Tab. A1 and Tab. A2. The accuracy of students can still be further degraded when learning from the stingy teacher on CIFAR-10 and Tiny-ImageNet.

A2.2 VISUALIZATION OF LOGITS

Figure A1 compares the soft probabilities produced by the normal teacher, the nasty teacher, and the stingy teacher respectively. Compared with the normal response, the nasty logit is very noisy and it significantly increases the probabilites of some irrelevant classes. As a result, it is easy to be identified by the attacker. Oppositely, the stingy logits still maintain the relatively relationships among the top categories, so it still provides the normal function as the original network.

Table A1: Comparison of the nasty teacher and the stingy teacher on CIFAR-10.

Teacher network	Teacher accuracy	Students accuracy after KD			
		CNN	ResNetC-20	ResNetC-32	ResNet-18
Student baseline	-	86.64	92.28	93.04	95.13
ResNet-18 (normal)	95.13	87.75 (+1.11)	92.49 (+0.21)	93.31 (+0.27)	95.39 (+0.26)
ResNet-18 (nasty)	94.56 (-0.57)	71.83 (**-14.81**)	74.22 (-18.06)	79.66 (-13.38)	91.55 (**-3.58**)
ResNet-18 (stingy)	95.13 (**-0.0**)	82.77 (-3.87)	68.86 (**-25.42**)	74.34 (**-18.70**)	92.46 (-2.67)

Table A2: Comparison of the nasty teacher and the stingy teacher on Tiny-ImageNet

Teacher network	Teacher accuracy	Students accuracy after KD			
		Shufflenetv2	MobilenetV2	ResNet-18	Teacher Self
Student baseline	-	55.74	51.72	58.73	-
ResNet-18 (normal)	58.73	58.09 (+2.35)	55.99 (+4.27)	61.45 (+2.72)	61.45 (+2.72)
ResNet-18 (nasty)	57.77 (-0.96)	23.16 (**-32.58**)	1.82 (**-49.90**)	44.73 (-14.00)	44.73 (-14.00)
ResNet-18 (stingy)	58.73 (**-0.00**)	34.36 (-21.38)	5.55 (-46.17)	33.34 (**-25.39**)	33.34 (**-25.39**)
ResNet-50 (normal)	62.01	58.01 (+2.27)	54.18 (+2.46)	62.01 (+3.28)	63.91 (+1.90)
ResNet-50 (nasty)	60.06 (-1.95)	41.84 (-13.90)	1.41 (**-50.31**)	48.24 (-10.49)	51.27 (-10.74)
ResNet-50 (stingy)	62.01 (**-0.00**)	28.03 (**-27.71**)	5.41 (-46.31)	37.05 (**-21.68**)	34.26 (**-27.75**)
ResNeXt-29 (normal)	62.81	57.87 (+2.13)	54.34 (+2.62)	62.38 (+3.65)	64.22 (+1.41)
ResNeXt29 (nasty)	60.21 (-2.60)	42.73 (-13.01)	1.09 (**-50.63**)	54.53 (-4.20)	59.54 (-3.27)
ResNeXt29 (stingy)	62.81 (**-0.00**)	30.98 (**-24.76**)	9.65 (-42.07)	30.70 (**-28.03**)	34.67 (**-28.14**)

A2.3 Results on ImageNet

The accuracy of the nasty teacher is sensitive to the adversarial weights ω in the retraining process, thus the model owner need to pay lots of effort to exploring the best ω. On the contrary, without the retraining process, our stingy teacher can be easily scaled up to huge datasets. We evaluate the stingy teacher on ImageNet. As shows in Table A3, when distilling from a stingy DenseNet-121, the accuracy of students can be degraded up to 37.55%. As a result, our stingy teacher is also more favorable for protecting huge models in real applications.

Table A3: Experimental results on ImageNet.

Model	Baseline	self-KD	KD (normal T)	KD (stingy T)
ResNet-18	69.84	70.42 (+0.58)	70.40 (+0.56)	32.29 (**-37.55**)

A2.4 Ablation Studies of Stingy Teacher on standard KD

Sample of subset The stingy teacher preserves the logits of the top N categories to maintain the dark knowledge. We denote it as *"top logits"* for short. We then explore the performance of other possibility to build the subset $\mathbf{M}$. Specifically, we design another sparse logits that concatenates the top one logits and the N - 1 smallest logits, and we denote it as *"least logits"* for short. The least logits can be regarded as the worst sparse logits, as it masks out all meaningful dark knowledge, and enlarges the least related categories instead. Figure A2 presents the comparison results. Obviously, at around the 10% sparse ratio, both top logits and least logits achieve the greatest damage to the student networks. This is consistent with our analysis in Eq. 5 that when r is small, the sparse logits should be able to degrade the student networks, whatever subset we use. When r is equal to $\frac{1}{K}$, both of them degenerate into the hard label, thus they have the same performance. When r is large, the damage from both types of logits is alleviated, as the weight on the second term of Eq. 5 is reduced. However, the least logits always leads to a worse accuracy of students than the top logits. We believe that the irrelevant classes in $\mathbf{M}$ provide harmful interference to the learning of students, and make the learning much difficult. This also reveals that dark knowledge is beneficial

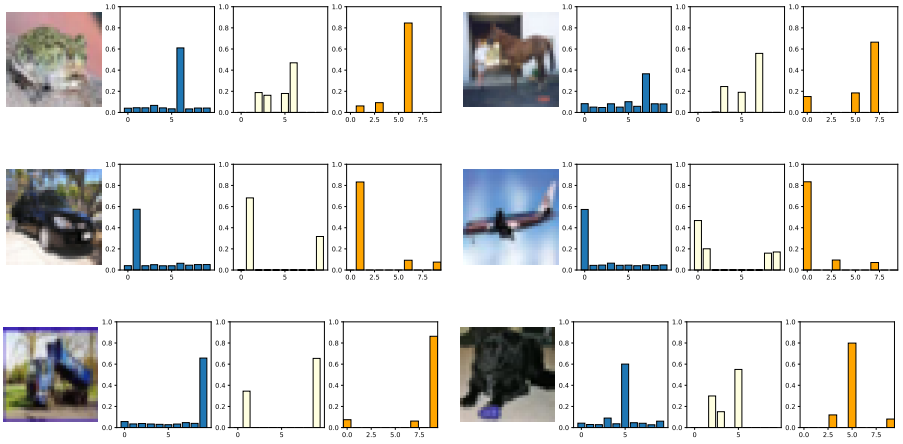

Figure A1: The visualization of logit responses produced by a normal ResNet-18 (blue), a nasty ResNet-18 (yellow), and a stingy ResNet-18 (orange) trained on CIFAR-10. We present the probabilities after temperature-scaled softmax, where τ is 4.

to the student networks. Although the performance of the least logits is promising, considering the similarity to the original logits, the top logits is still a favorable choice of the stingy teacher.

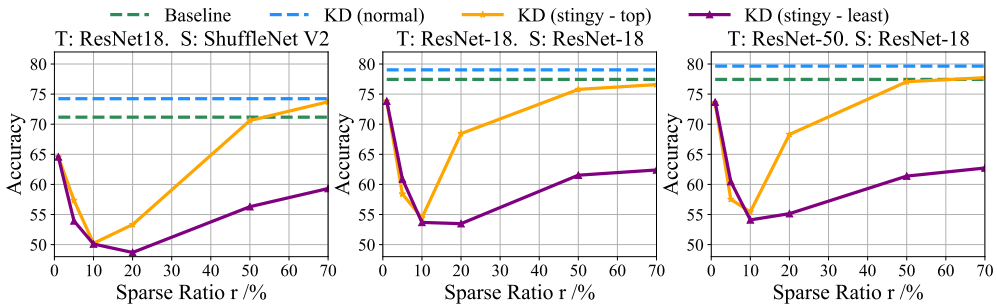

Figure A2: Comparison of sparse logits built with top N categories (top logits) and the combination of the top-1 class and N-1 smallest probabilities (least logits). Experiments are conducted on CIFAR-100.

Temperature We also conduct ablation studies to explore the effect of τ on the stingy teacher. We keep the sparse ratio $r = 0.1$ and vary the temperature τ_s from 1 to 20. As in Figure A3, with a larger τ, the student can also be further degraded when learning from the stingy teacher. When reducing τ, the student networks can recover some performance. As supported by Equation 5, a small τ turns the weights of the second term down, and thus mitigates its negative effect. Moreover, we cannot approximate the soft probabilities $p_\tau^T(k)$ with the uniform distribution when τ is small.

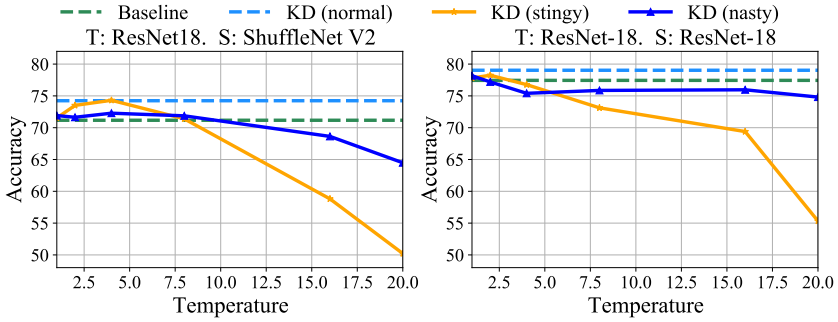

Figure A3: Ablation studies on temperature τ. Experiments are conducted on CIFAR-100.