南京航空航天大學

Nanjing University of Aeronautics and Astronautics

ParNeC 模式识别与神经计算研究组

PAttern Recognition and NEural Computing

Undistillable:

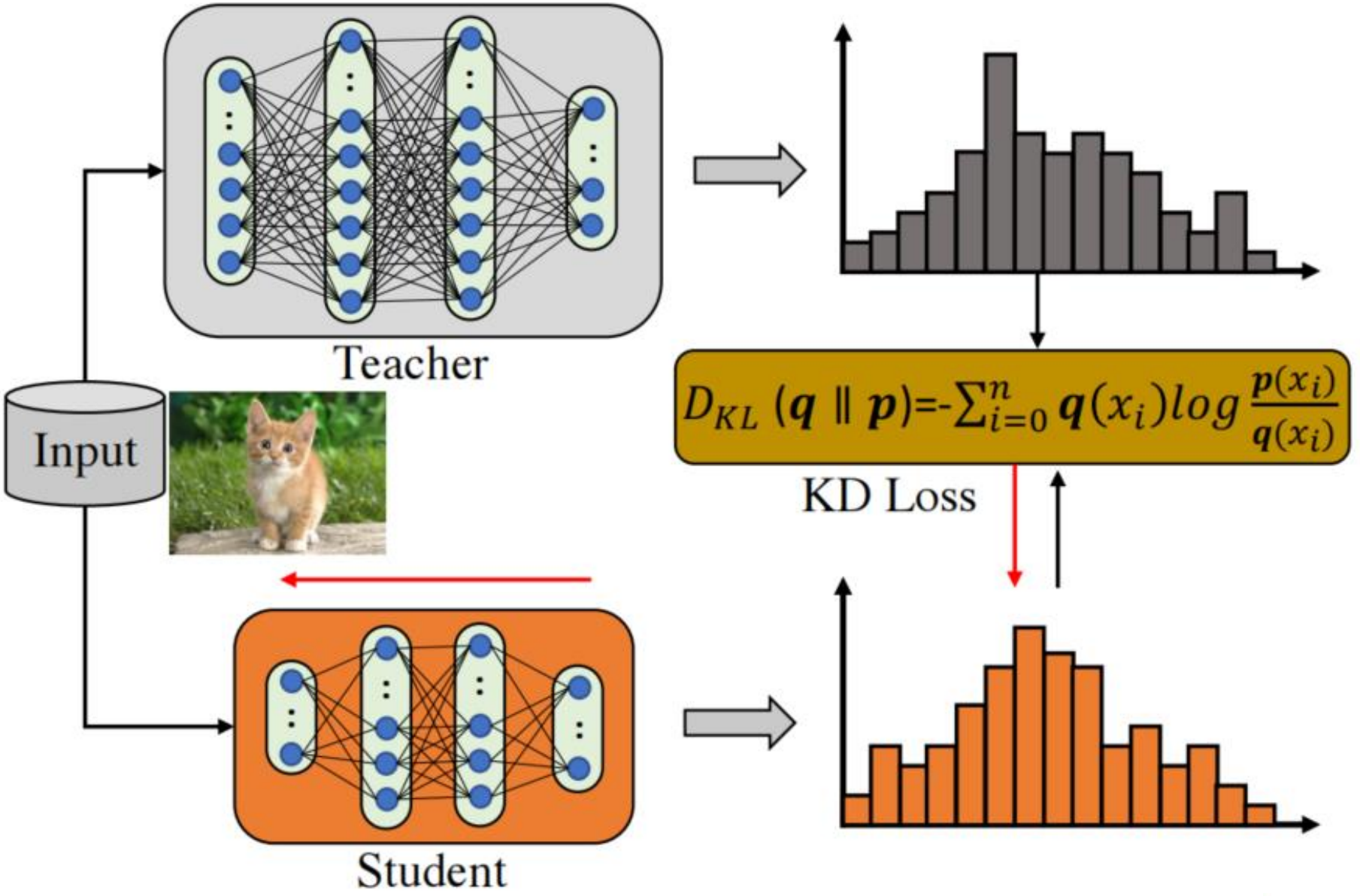
Making A Nasty Teacher that CANNOT Teach Students

Haoyu Ma[1], Tianlong Chen[2], Ting-Kuei Hu[3], Chenyu You[4], Xiaohui Xie[1], Zhangyang Wang[2]

[1]University of California, Irvine, [2]University of Texas at Austin, [3]Texas A&M University, [4]Yale University

ICLR 2021

Knowledge Distillation



Knowledge Distillation

$f_{\theta_T}(\cdot)$: a pre-trained teacher network

$f_{\theta_S}(\cdot)$: a student network

(x_i, y_i): a training sample in dataset $\mathcal{X}$

$p_{f_\theta}(x_i)$: the logit response of x_i from $f_\theta(\cdot)$

$\mathcal{KL}(.)$: K-L divergence

$\mathcal{XE}(\cdot,\cdot)$: cross-entropy loss

$\sigma_{\tau_s}(\cdot)$: softmax temperature function

the student network f_{θ_S} could be learned by the following:

$$\min_{\theta_S} \sum_{(x_i,y_i)\in\mathcal{X}} \alpha\tau_s^2\mathcal{KL}(\sigma_{\tau_s}(p_{f_{\theta_T}}(x_i)), \sigma_{\tau_s}(p_{f_{\theta_S}}(x_i))) + (1-\alpha)\mathcal{XE}(\sigma(p_{f_{\theta_S}}(x_i)), y_i)$$

Training Nasty Teachers

$f_{\theta_T}(\cdot)$: the desired nasty teacher

$f_{\theta_A}(\cdot)$: its adversarial learning counterpart

(pre-trained in advance and fixed)

τ_A denotes the temperature for self-undermining

ω balances the behavior between normal training and adversarial learning

$$\min_{\theta_T} \sum_{(x_i,y_i)\in\mathcal{X}} \mathcal{XE}(\sigma(p_{f_{\theta_T}}(x_i)), y_i) - \omega\tau_A^2 \mathcal{KL}(\sigma_{\tau_A}(p_{f_{\theta_T}}(x_i)), \sigma_{\tau_A}(p_{f_{\theta_A}}(x_i)))$$

Experiments

Table 1: Experimental results on CIFAR-10.

Teacher network	Teacher performance	Students performance 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)	82.46 (-4.18)	88.01 (-4.27)	89.69 (-3.35)	93.41 (-1.72)

Table 2: Experimental results on CIFAR-100.

Teacher network	Teacher performance	Students performance 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-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)
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)

Experiments

Table 3: Experimental results on Tiny-ImageNet

Teacher network	Teacher performance	Students performance 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-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)
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)

Qualitative Analysis

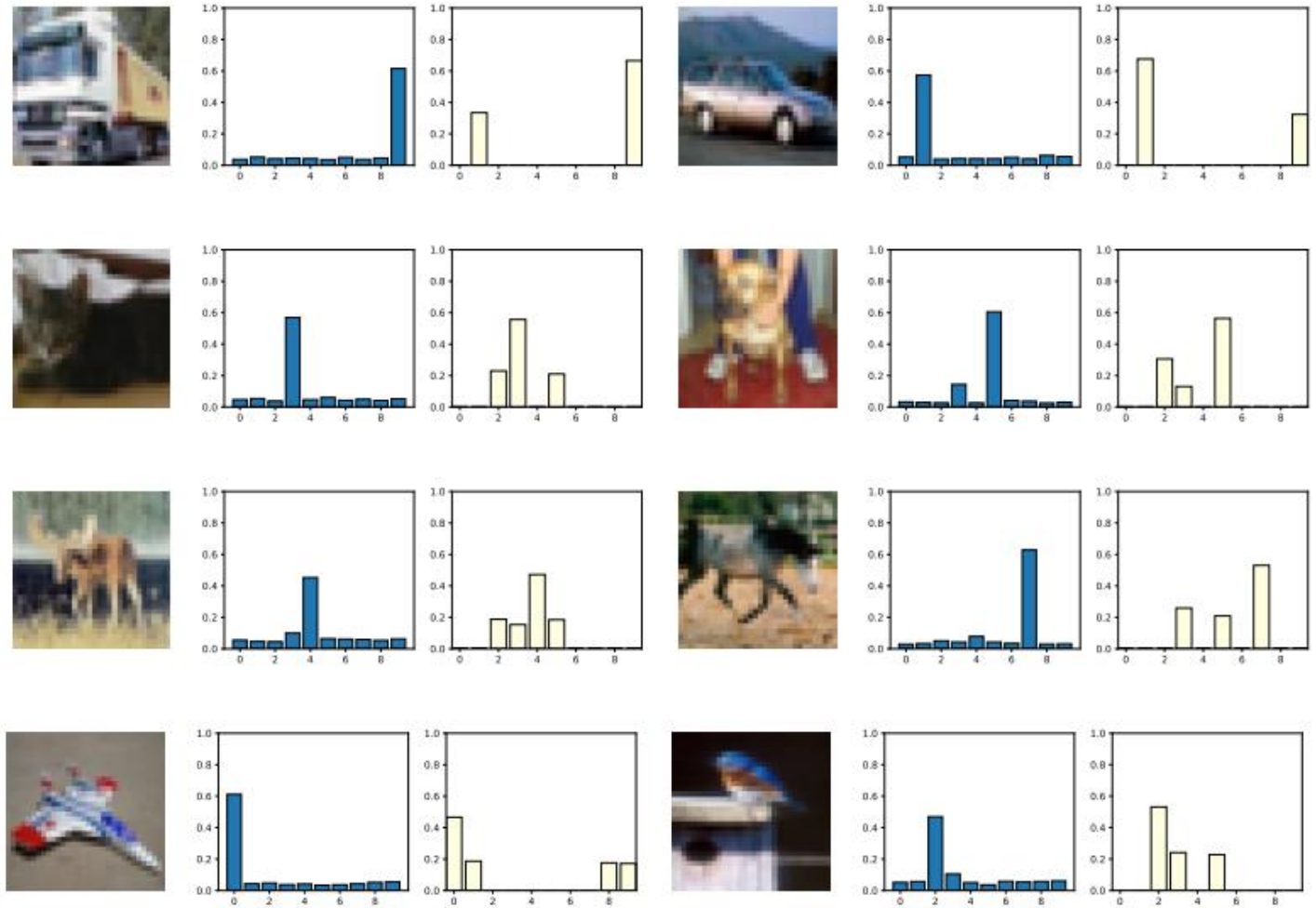
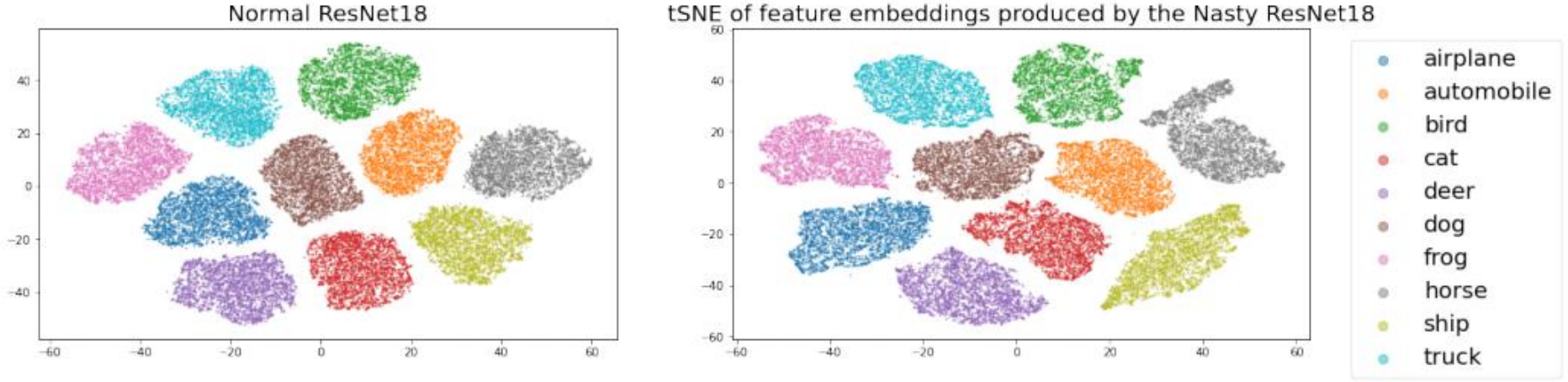
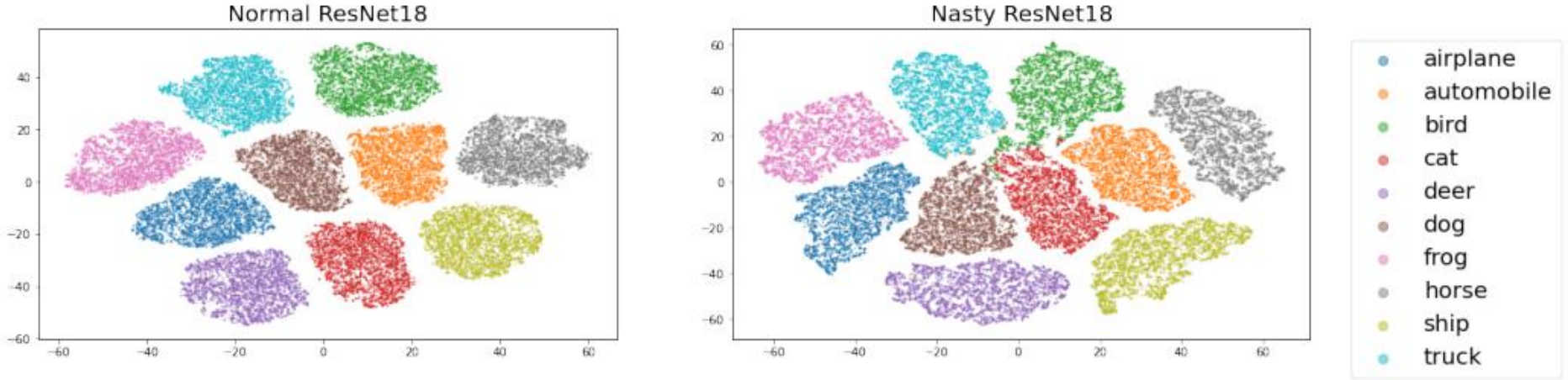


Figure 1: The visualization of logit responses after "temperature softmax" function. Each row represents two examples from CIFAR-10. The sampled images are shown in the 1st and the 4th columns. The 2nd and the 5th columns summarize the scaled output from the normal teacher. The 3rd and the 6th columns represent the scaled output from the nasty teacher

Qualitative Analysis



(a) tSNE of feature embeddings before fully-connected layer. The dimension of feature embeddings is 512.



(b) tSNE of output logits

Figure 2: Visualization of tSNEs for both normal and nasty ResNet18 on CIFAR-10. Each dot represents one data point.

Ablation Study: Adversarial Network

1. weak networks (e.g., Plain CNN) might lead to less effective nasty teachers.
2. although stronger networks contribute to more effective nasty teachers, we observe that the trade-off accuracy is saturated quickly and converges to "self-undermining" ones.

Table 4: Ablation study w.r.t the architecture of the adversarial network $f_{\theta_A}(\cdot)$ on CIFAR-10.

Teacher network	Teacher performance	Students after KD			
		CNN	ResNetC20	ResNetC32	ResNet18
Student baseline	-	86.64	92.28	93.04	95.13
ResNet18(normal)	95.13	87.75 (+1.11)	92.49 (+0.21)	93.31 (+0.27)	95.39 (+0.26)
ResNet18(ResNet18)	94.56 (-0.57)	82.46 (-4.18)	88.01 (-4.27)	89.69 (-3.35)	93.41 (-1.72)
ResNet18(CNN)	93.82 (-1.31)	77.12 (-9.52)	88.32 (-3.96)	90.40 (-2.64)	94.05 (-1.08)
ResNet18(ResNeXt-29)	94.55 (-0.58)	82.75 (-3.89)	88.17 (-4.11)	89.48 (-3.56)	93.75 (-1.38)

Ablation Study: Students Network

As the Reversed KD in Yuan et al. (2020), the superior network can also be enhanced by learning from a weak network. To explore the generalization ability of our method, we further conduct experiments on the reversed KD.

Table 5: Ablation study w.r.t the architecture of the student networks.

Dataset	CIFAR-10		CIFAR-100	
Student network	ResNet-50	ResNeXt-29	ResNet-50	ResNeXt-29
Student baseline	94.98	95.60	78.12	81.85
KD from ResNet-18 (normal)	94.45 (-0.53)	95.92 (+0.32)	79.94 (+1.82)	82.14 (+0.29)
KD from ResNet-18 (nasty)	93.13 (-1.85)	92.20 (-3.40)	74.28 (-3.84)	78.88 (-2.97)

Ablation Study

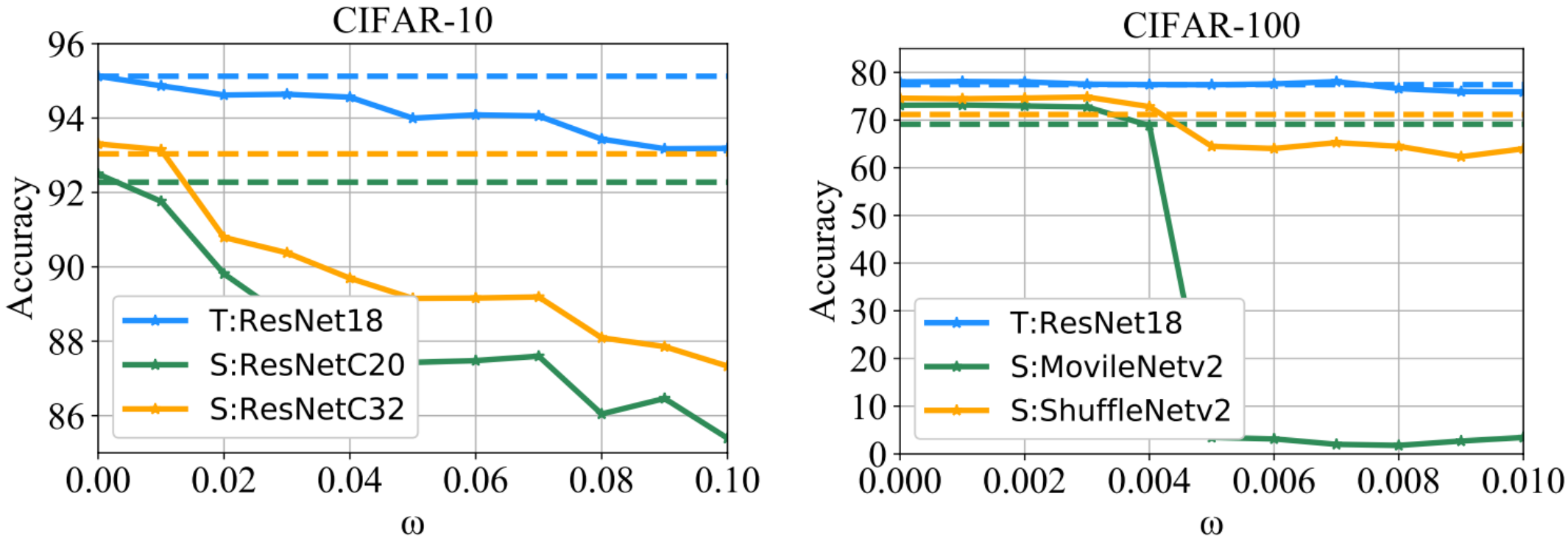
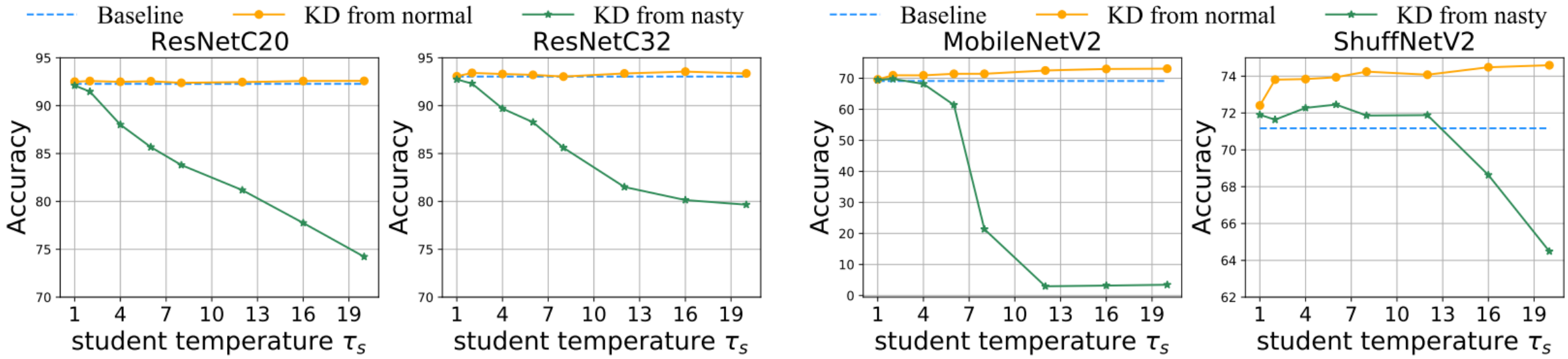


Figure 3: Ablation study w.r.t ω on CIFAR-10 and CIFAR-100. The initials "T" and "S" in the legend represent teacher networks and student networks, respectively. The dash-line represents the accuracy that the model is normally trained.

Ablation Study



(a) Nasty teacher with $\tau_A = 4$ on CIFAR-10

(b) Nasty teacher with $\tau_A = 20$ on CIFAR-100

Figure 4: Ablation study w.r.t temperature τ_s. The architecture of teacher networks are ResNet-18 for both CIFAR-10 and CIFAR-100 experiments. Each figure presents accuracy curves of student networks under the guidance of the nasty or normal ResNet-18 with various temperature τ_s.

Nasty Teacher on Data-Free KD

Table 6: 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)

(a) Normal Teacher

(b) Nasty Teacher

Figure 6: Images generated by inverting a normal ResNet34 and a nasty ResNet34 trained on CIFAR-10 with DeepInversion. For each image, each column represents one category.

ICLR2022: Non-Transferable Learning

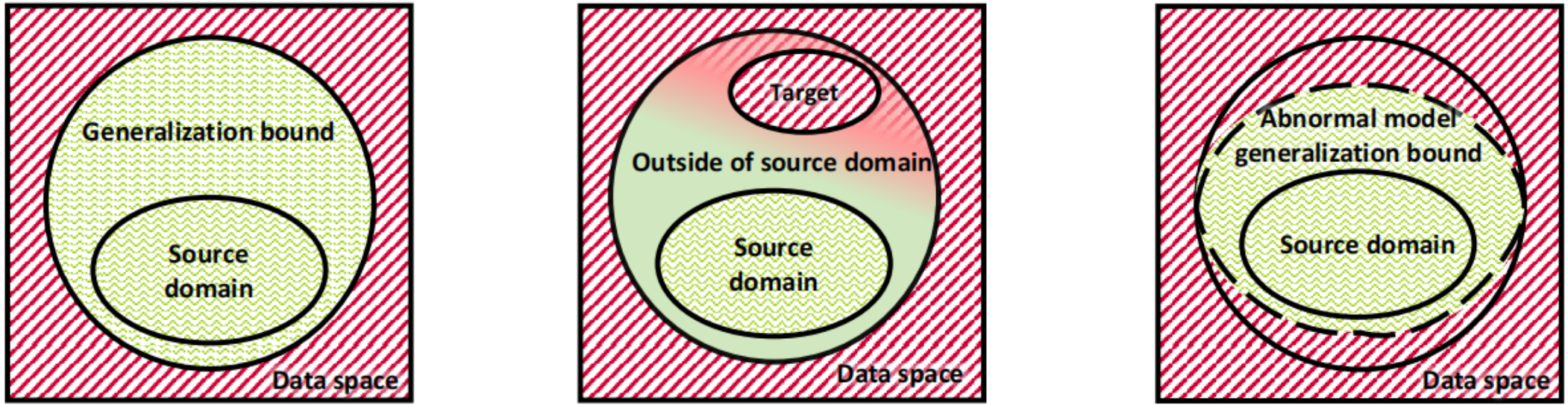


Figure 1: **A visualization of the generalization bound trained with different approaches.** The left figure shows Supervised Learning in the source domain, which can derive a wide generalization area. When Target-Specified NTL is utilized (middle), the target domain is removed from the generalization area. As for Source-Only NTL (right), the generalization area is significantly reduced.