# Undistillable:
# Making a Nasty Teacher that Cannot Teach Students

Haoyu Ma, Tianlong Chen, Ting-Kuei Hu, Chenyu You, Xiaohui Xie, Zhangyang Wang(UC Irvine)
ICLR 2021.

**Presenter: Dongyu Yao**

# Background

## Knowledge Distillation[1] -- Model Compression

- Transfer useful knowledge from Teacher Network(High accuracy, complex) to Student Network(High accuracy, simple)

- Data Driven KD: Train the Student Network on the same dataset as the Teacher

- Data Free KD: Student Network has no access to the orginal dataset

## Intellectual Property Infringement

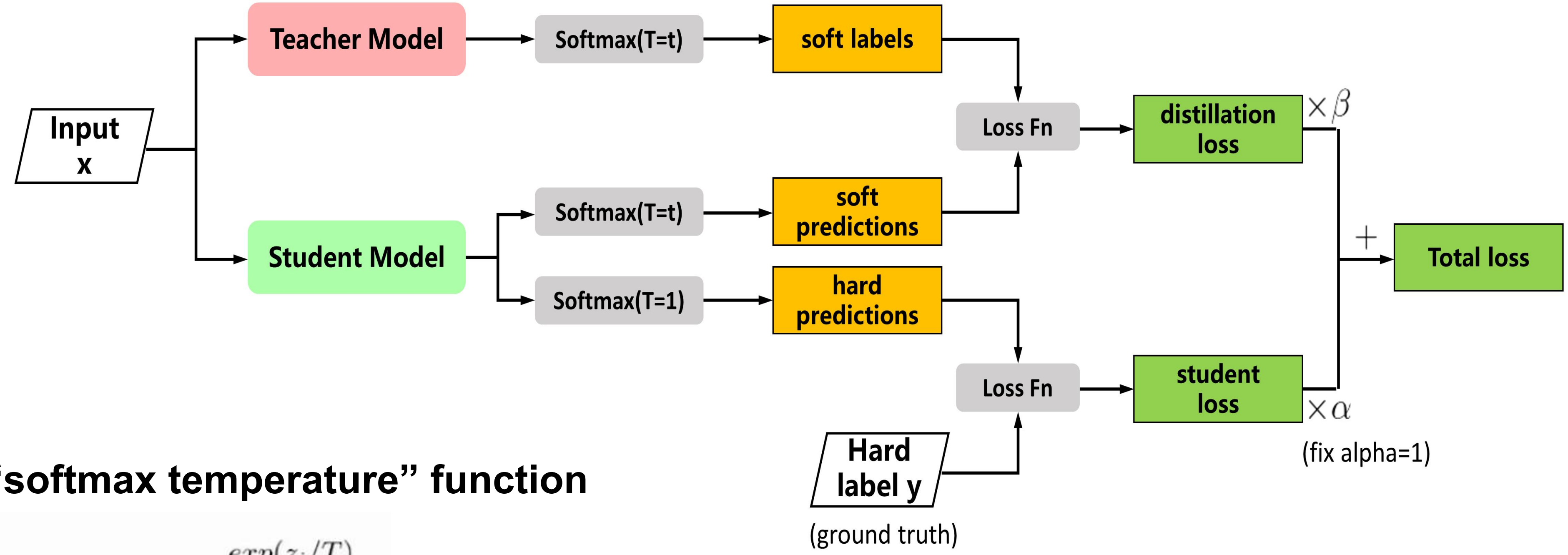
- Data Driven KD: Easily stealing the well trained model

- Data Free KD: Restoring potential personal training dataset, threating the owner’s data privacy and security

[1] Geoffrey Hinton, et al. Distilling the Knowledge in a Neural Network. https://arxiv.org/abs/1503.02531

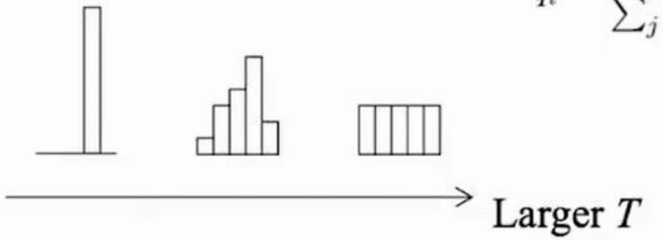
# Related Work

## Knowledge Distillation




**“softmax temperature” function**

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$




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

Images from: https://nni.readthedocs.io/en/v2.5/TrialExample/KDExample.html

# Motivation

## How to protect the model?

- Train an Undistillable Teacher Model--Nasty Teacher:
keep the performance when normally using, but deprecate the performance when distilled into Student Model

## Different from Model Watermarking:

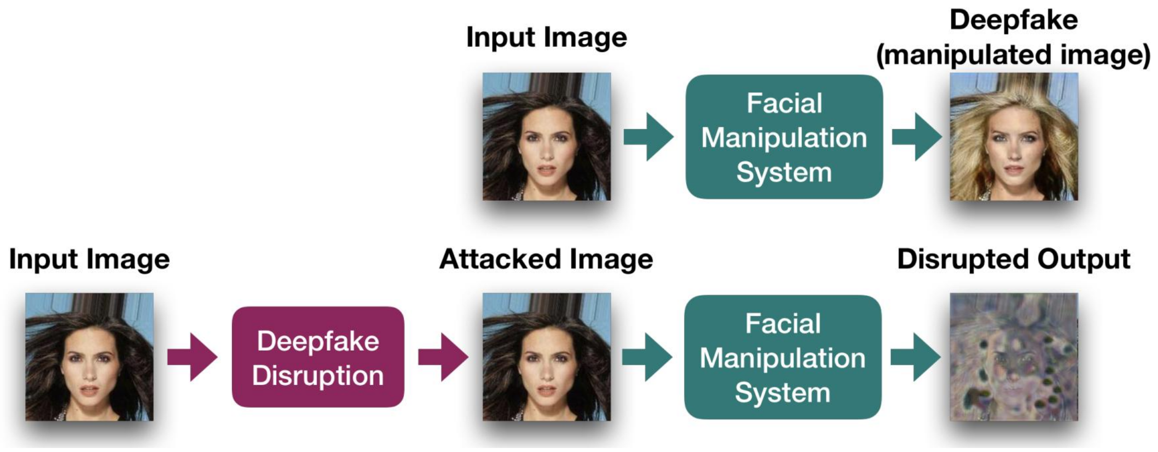
- Watermarking: “Afterwards detection”, verifying the ownership via water mark **after stolen**

- Nasty Teacher: “**Proactive defence**”, making the stolen model unavailable

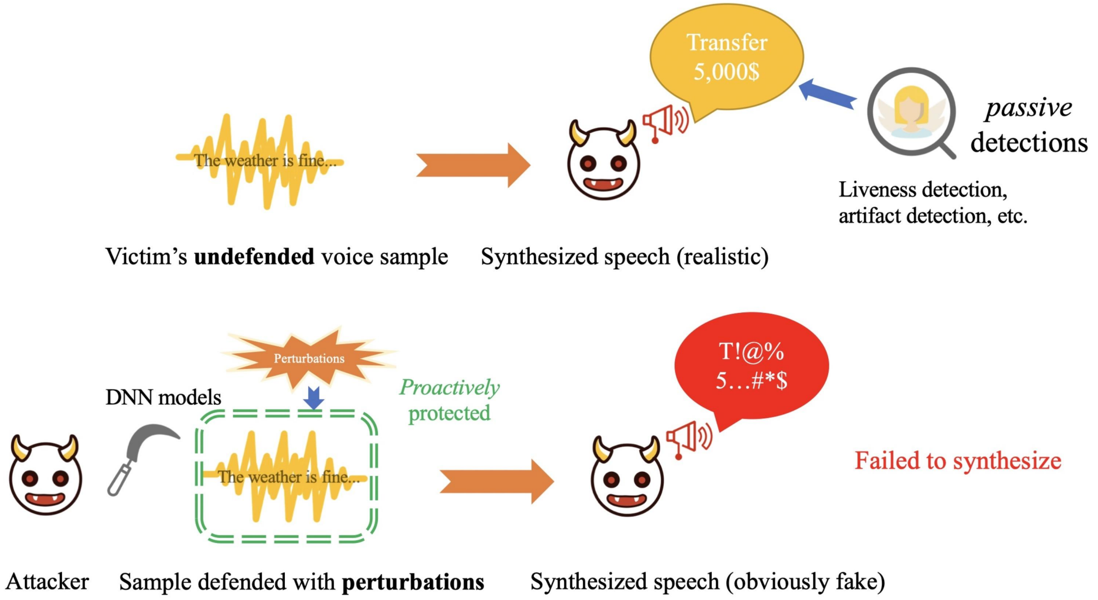
# Does “proactive” sound familiar?

**Yes! We have cases on images[2] and voices[3]!**

- Disrupting Deekfakes



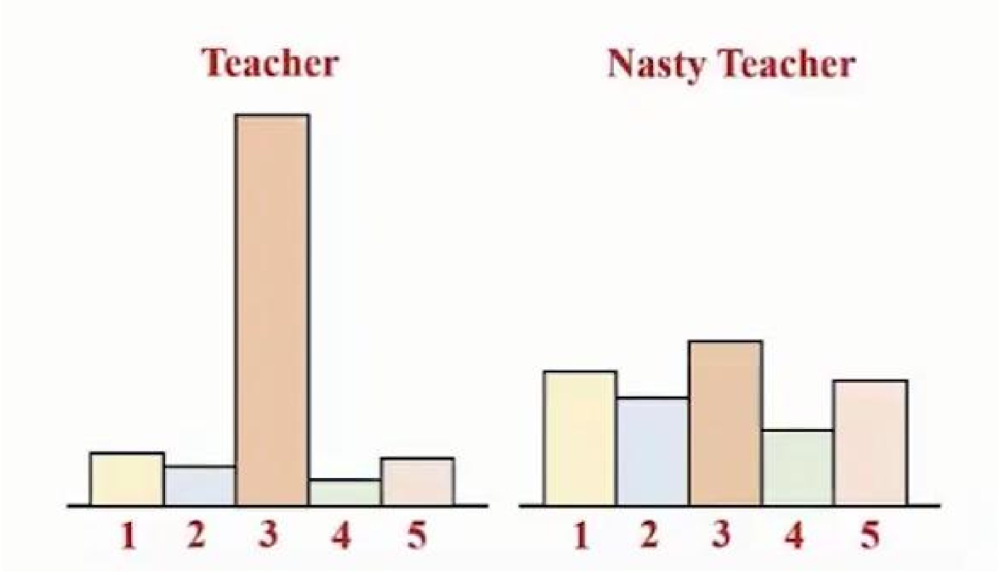

- Defending your voice




[2] Ruiz, N et al. 2020. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In ECCV.
[3] Huang et al. 2020. Defending Your Voice: Adversarial Attack on Voice Conversion https://arxiv.org/abs/2005.08781

# Methodology

- Propose Self-Undermining Knowledge Distillation




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

$\theta_A$: Pretrained parameters(Fix temperature T)
$\theta_T$: Nasty teacher parameters(only ones to be updated)
$\sigma_{TA}$: "softmax temperature" function
XE: Cross-Entropy loss
KL: Kullback-Leibler divergence loss

# Experiment

## Results

CIFAR10 ($\tau_A$=4, $\omega$=0.004)，CIFAR100 ($\tau_A$=20, $\omega$=0.005)，Tiny-ImageNet ($\tau_A$=20, $\omega$=0.01)

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) |

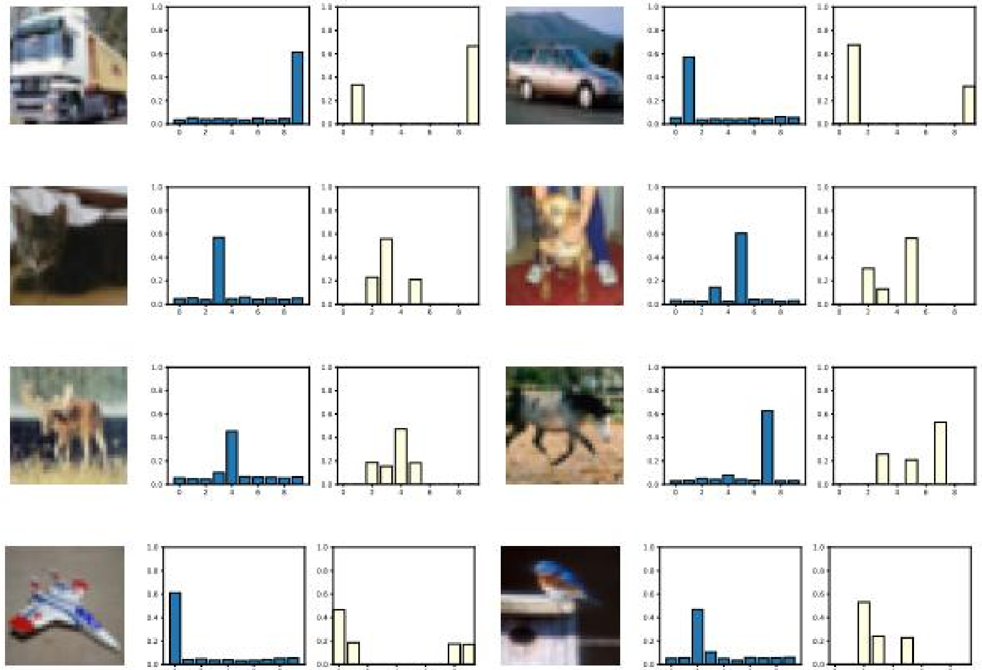
Table 3: Experimental results on Tiny-ImageNet

| 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) |

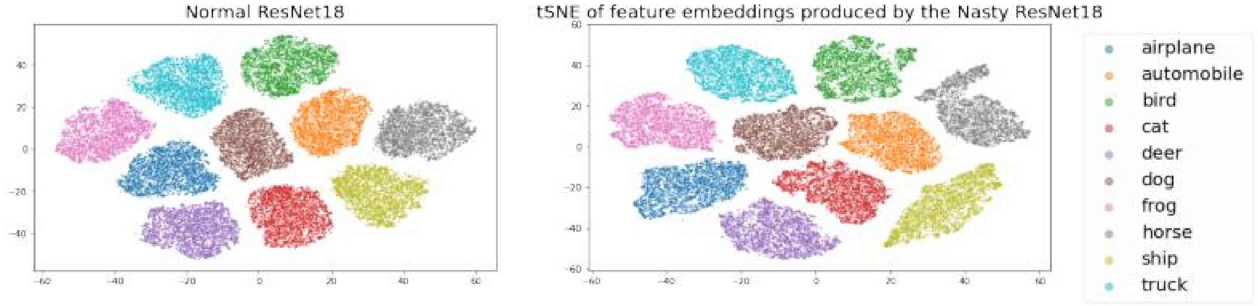
**Deeply deprecated**

# Experiment

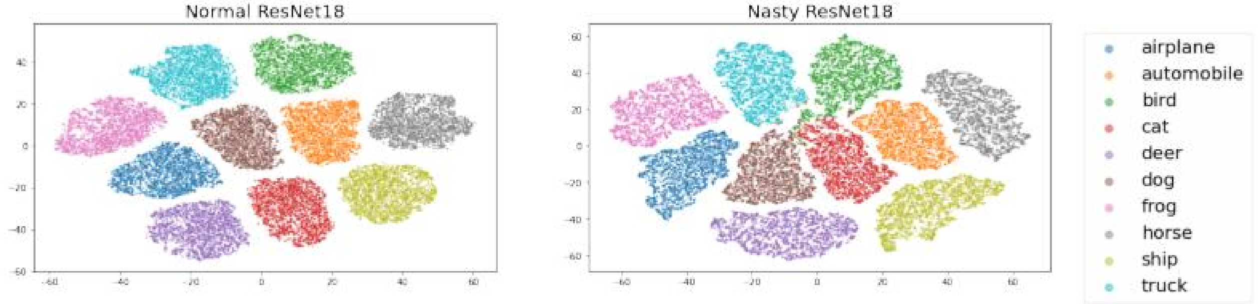
## Quantitative Analysis



The visualization of logit responses after ”temperature softmax” function.




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




(b) tSNE of output logits

# Experiment

## Ablation Study

### Adversarial Network

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) |

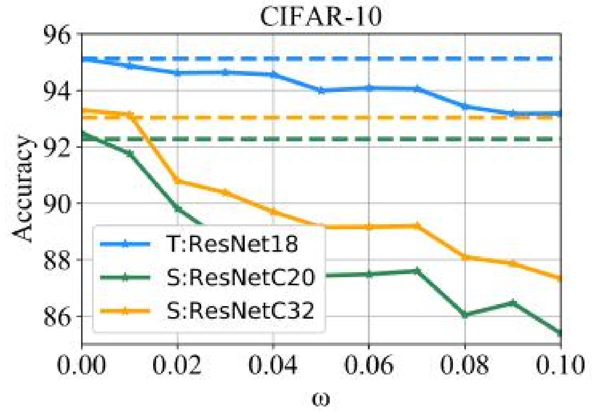
### Student Network

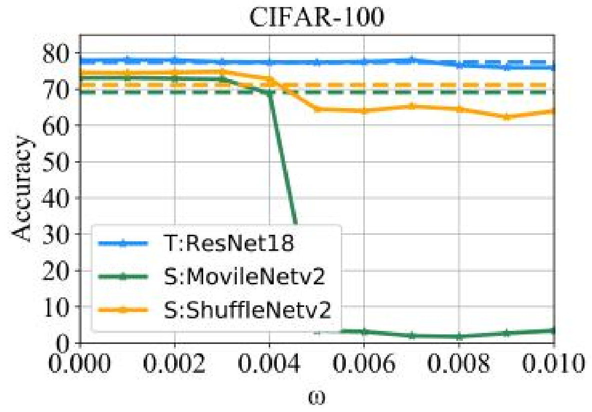
Table 5: Ablation study w.r.t the architecture of the student networks.

| 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) |

### Hyper parameters ω

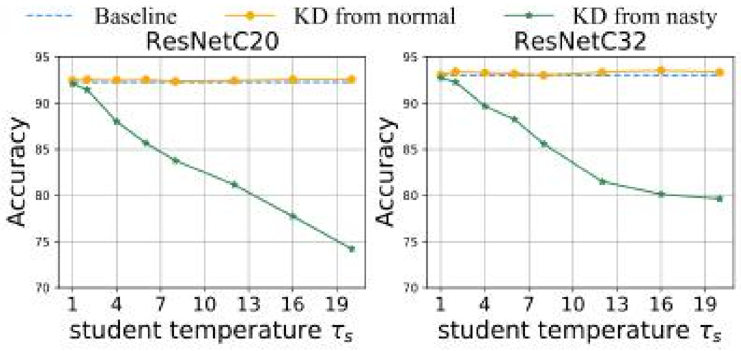
$$\min_{\theta_T} \sum_{(x_i,y_i)\in\mathcal{X}} \mathcal{XE}(\sigma(p_{f_{\theta_T}}(x_i)), y_i) - \omega\tau_A^2 \mathcal{KL}(\sigma_{\tau_A}(p_{f_{\theta_T}}(x_i)), \sigma_{\tau_A}(p_{f_{\theta_A}}(x_i))),$$



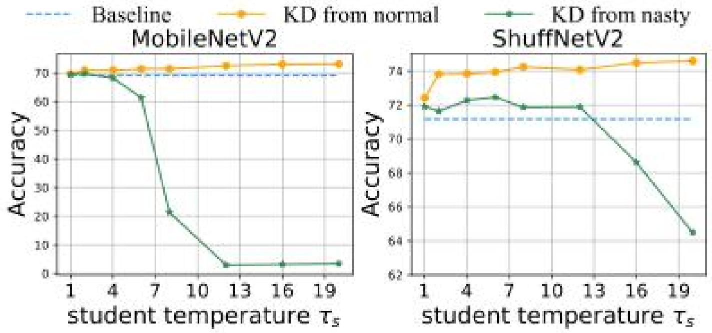



# Ablation Study

## Temperature $T_s$ In KD



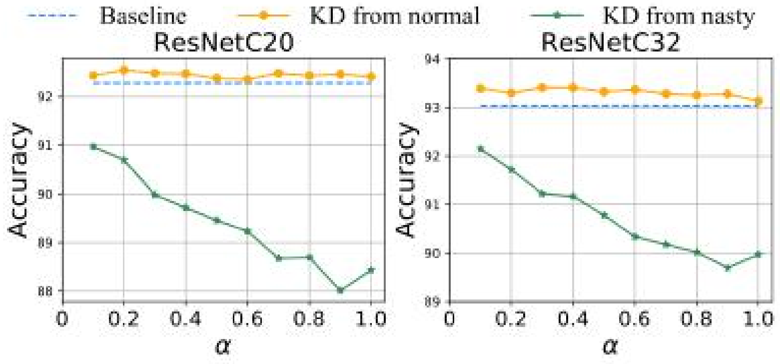

(a) Nasty teacher with $\tau_A = 4$ on CIFAR-10




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

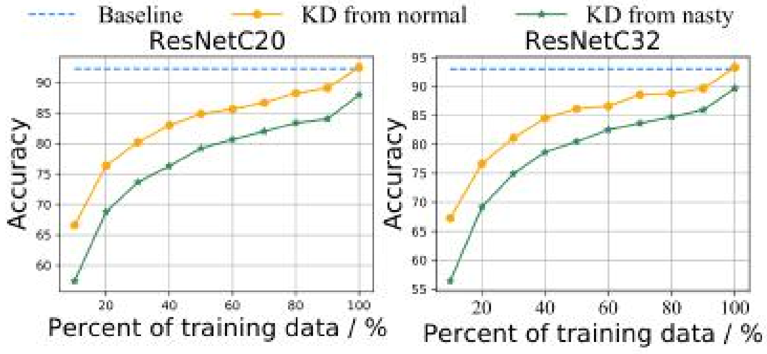
## KL Ratio α In KD

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




(a)

## Percentage of Training Samples




(b)

# NASTY TEACHER ON DATA-FREE KD

**DAFL**

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) |

**DeepInversion**

(a) Normal Teacher

(b) Nasty Teacher

## Some thoughts

· Can we simply add perturbations to the output logits instead of training an adversarial network?

· Transfer this idea into other areas： image processing; semantic segmentation

# Thank you