Undistillable: Making A Nasty Teacher That CANNOT teach students

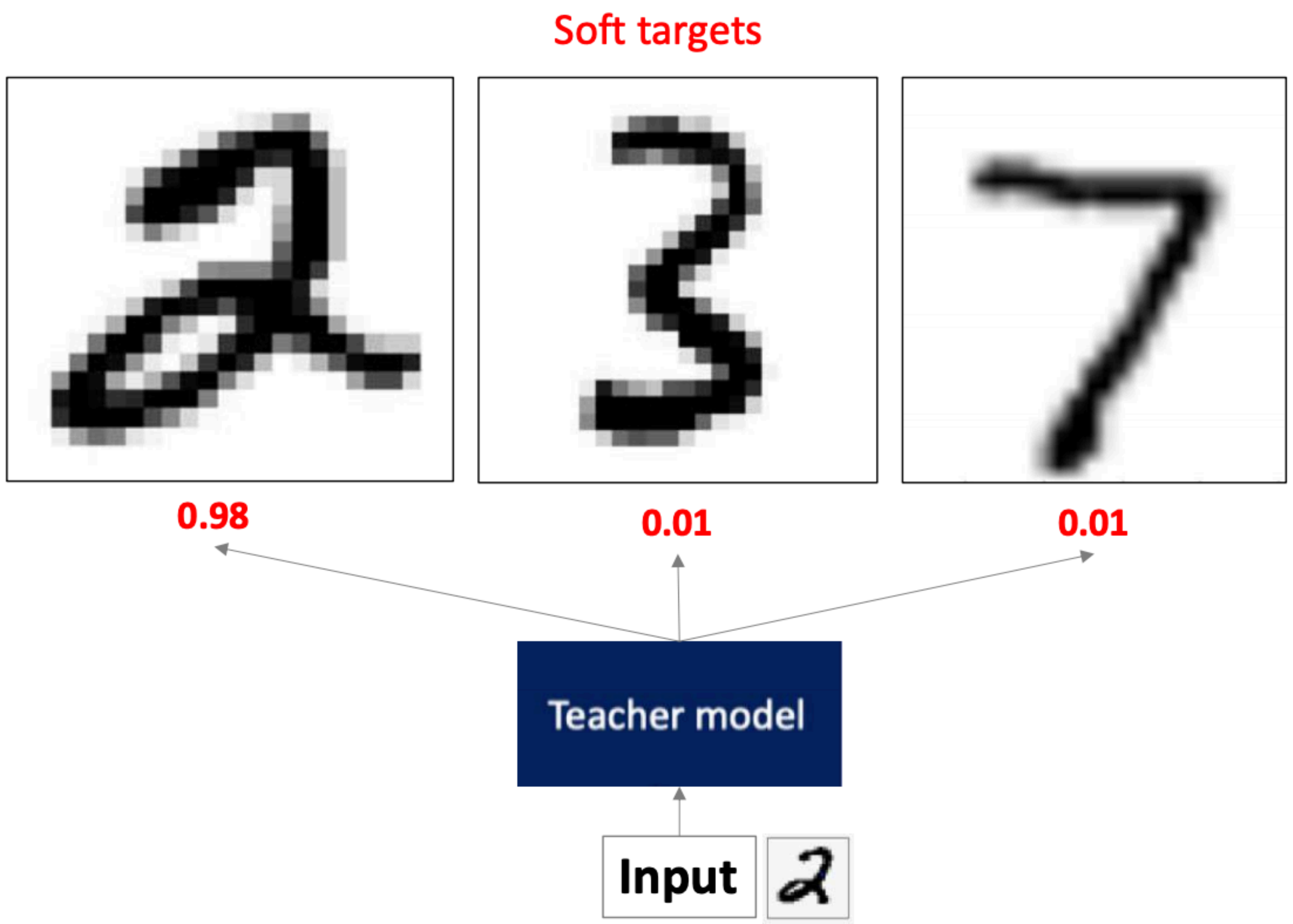
ICLR 2021 (Spotlight)

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

Background: Soft Target

2 is similar to 3 and 7

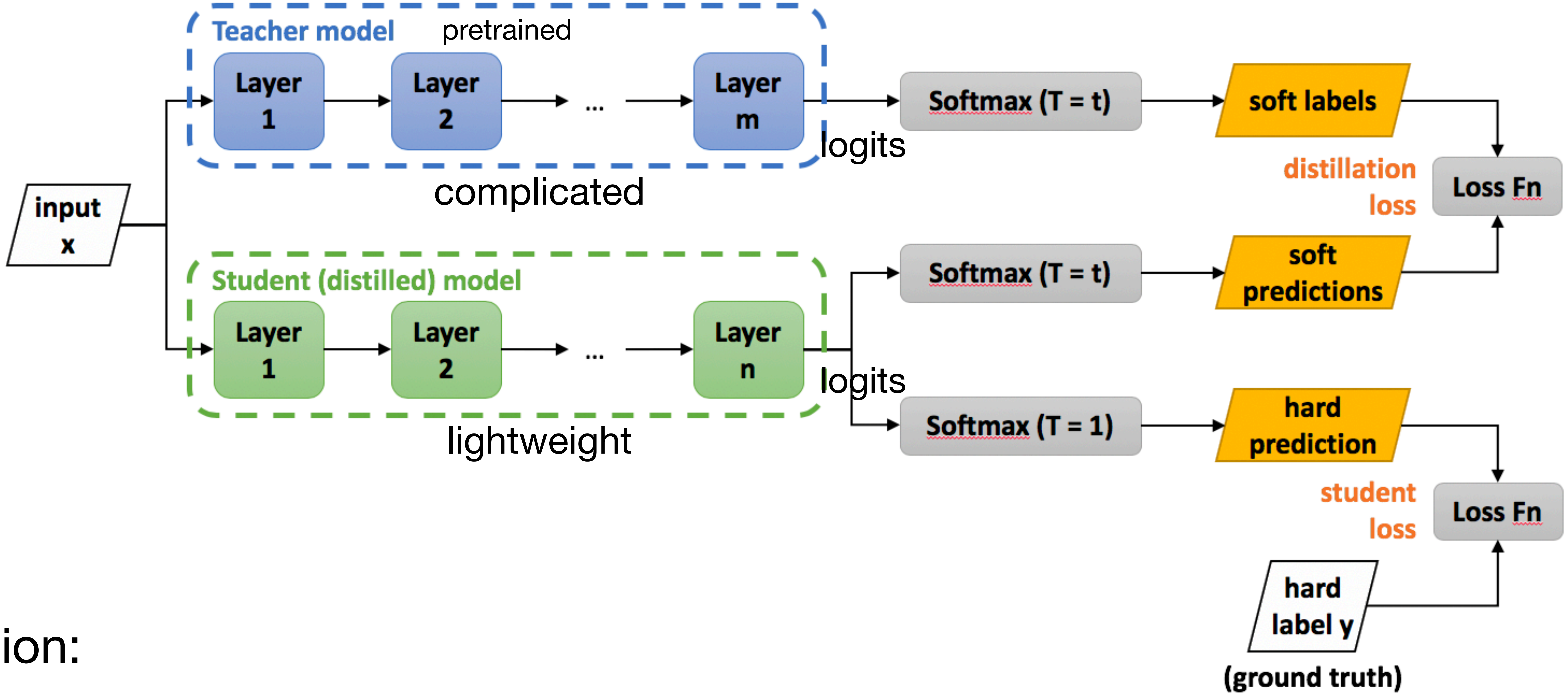
- Soft targets
  - Inter-class variance
  - Between-Class distance



Reference: http://ir.hit.edu.cn/~xiachongfeng/slides/Knowledge%20Distillation.pdf

Background: Knowledge Distillation (KD)

- Knowledge distillation framework

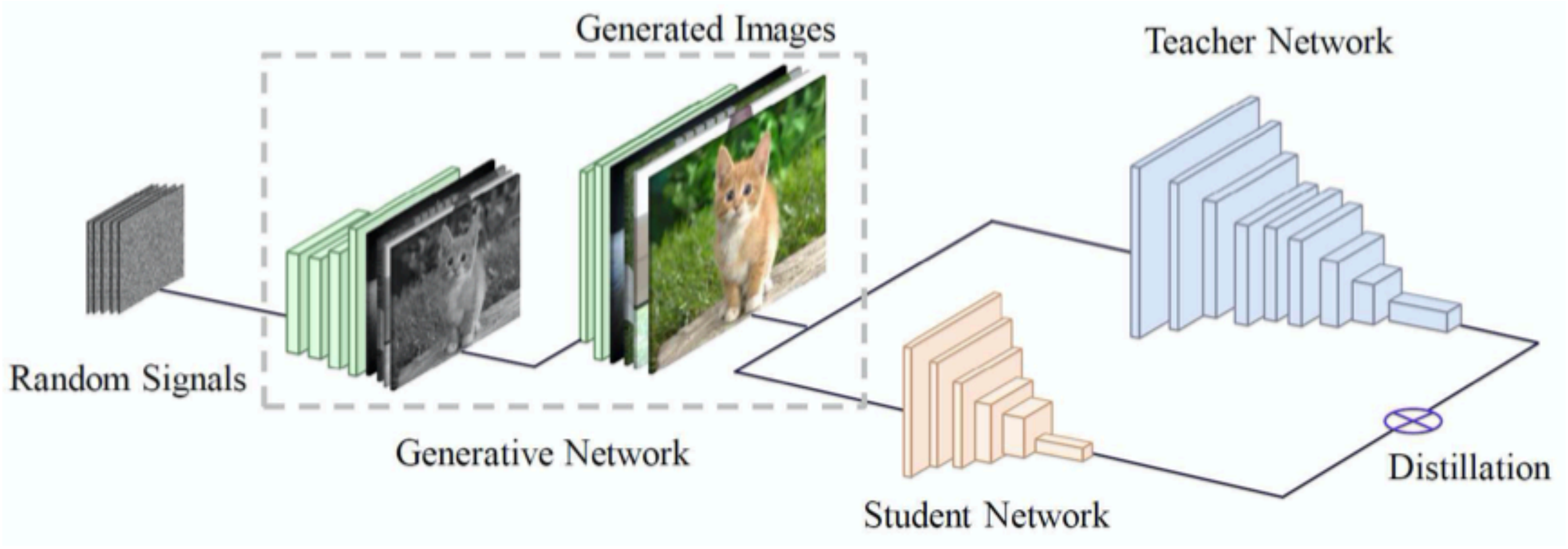


- Application:
    - Compression: high performance and lightweight student models

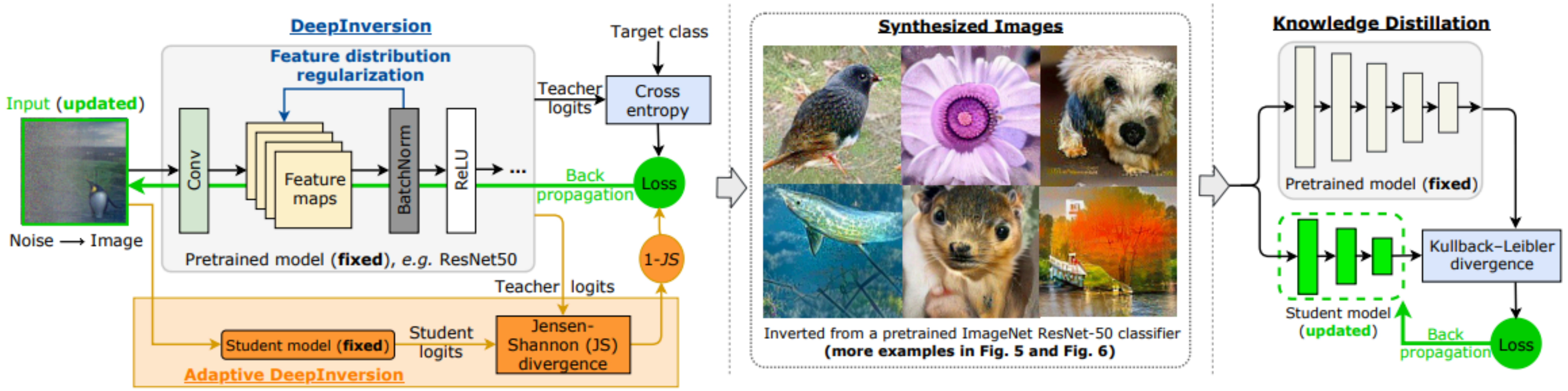
Reference: https://intellabs.github.io/distiller/knowledge_distillation.html

Background: Data-Free Knowledge Distillation

- Learn from the input-output behaviors without training data



DAFL: Chen, et al. Data-Free Learning of Student Networks, ICCV 2019



DeepInversion: Dreaming to Distill: Data-free Knowledge Transfer via DeepInversion, CVPR 2020

The unwanted side effects of KD

risk to the machine learning intellectual property (IP) protection

- 1) AI competition:
  - can obtain original training examples
  - clone the performance of top player's advanced networks from their logits (standard KD)
- 2) commercial black-box API:
  - cannot obtain original training examples
  - clone the functionality by imitating the input-output behaviors (data-free KD)

Nasty Teacher

A defensive approach to prevent knowledge leaking through KD

- Achieves nearly the same performance as its normal counterpart
- Significantly Degrades the performance of models that try to imitate it through KD

Methodology: Self-Undermining Knowledge Distillation

- Motivation: maintain correct class assignments, maximally disturbing incorrect class assignments so that no beneficial information could be distilled

- Implementation:
  - step 1: Train a normal teacher network $f_{\theta_A}(\cdot)$, named adversarial network
  - step 2: Train the nasty teacher (same architecture) $f_{\theta_T}(\cdot)$ by:

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

Results

Standard KD

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)

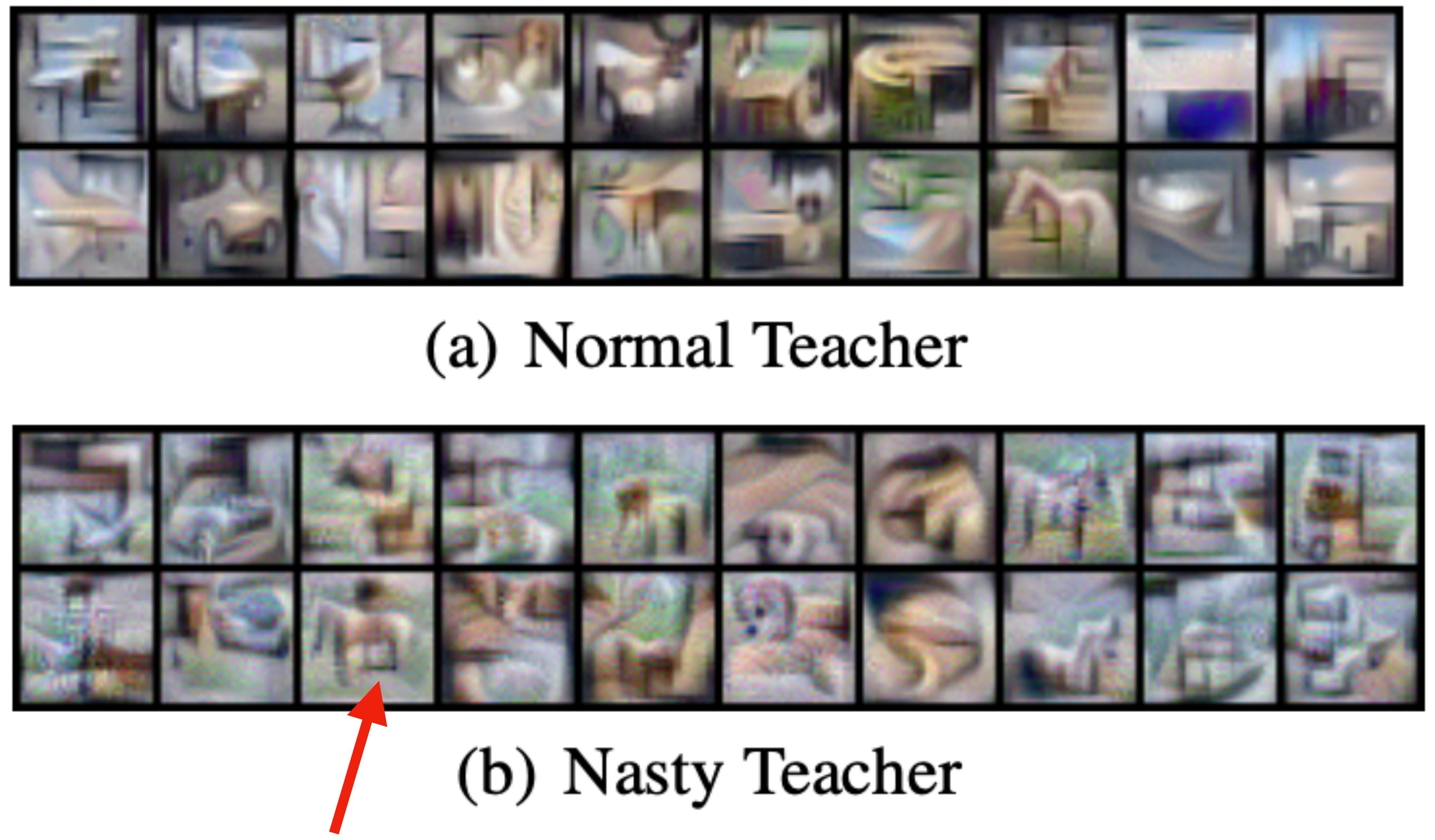
Data-Free KD

DAFL

Table 5: 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)

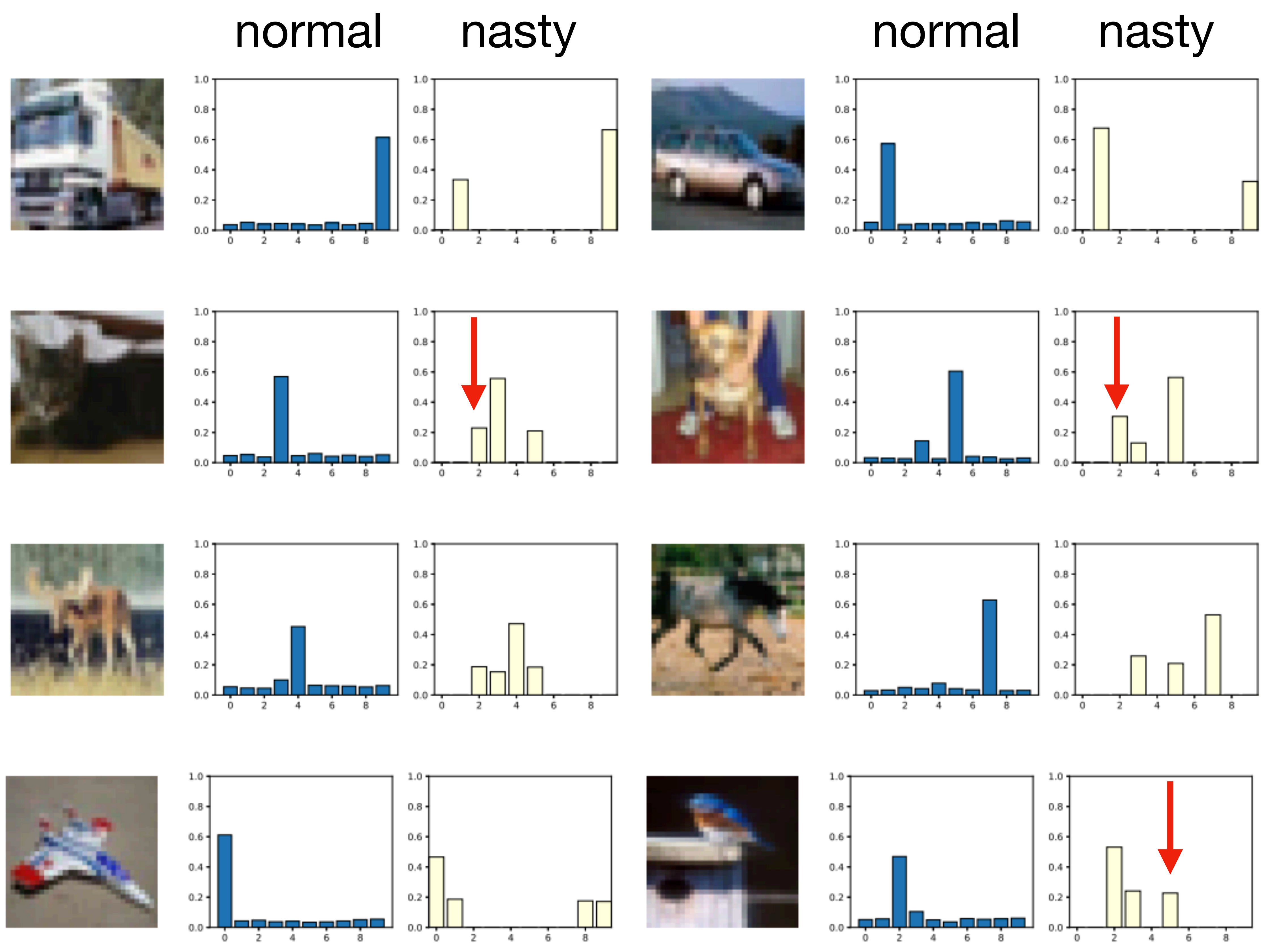
DeepInversion



(a) Normal Teacher

(b) Nasty Teacher

Analysis



- Logit response of nasty networks consists of sparse multiple peaks

- The multi-peak logits may give a false sense of generalization and thus mislead the learning of students

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

Code is available

Motivations

- Knowledge Distillation (KD) might open a loophole to unauthorized infringers to clone the Intellectual Property (IP) model's functionality.
- Data-Free KD [1][2] eliminates the necessity of accessing original training data, therefore can clone the functionality by simply imitating the input-output behavior.

Concept: Nasty Teacher

- A specially trained network that yields nearly the same performance as a normal one; but if used as a teacher model, it will significantly degrade the performance of student models that try to imitate it.

Methodology: Self-Undermining KD

- Rationale: maintain correct class assignments, maximally disturbing incorrect class assignments so that no beneficial information could be distilled.

- Implementation:
  1) Train a normal teacher network (adversarial network)
  2) Train the nasty teacher by maximizing the K-L divergence between the output of the nasty teacher and the adversarial network and simultaneously minimizing the cross entropy loss with the label.

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

References

[1] Chen, Hanting, et al. "Data-free learning of student networks." ICCV 2019.
[2] Yin, Hongxu, et al. "Dreaming to distill: Data-free knowledge transfer via deepinversion." CVPR 2020

Results

Standard KD

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.

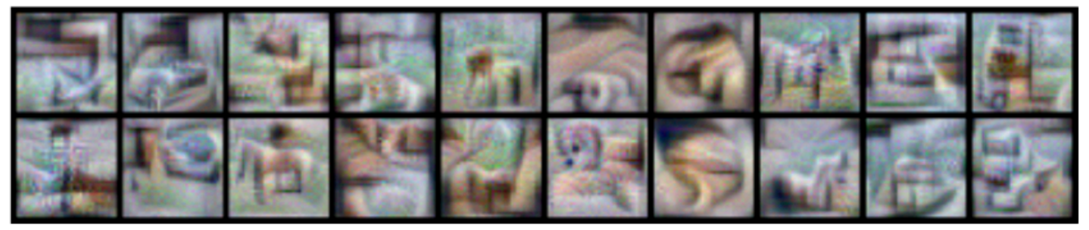
Teacher network	Teacher performance	Students performance after KD			
		Shufflenetv2	MobilenetV2	ResNet-18	Teacher Self
Student baseline	-	71.17	69.12	77.44	-
ResNet-18 (normal)	77.44	74.24 (+3.07)	73.11 (+3.99)	79.03 (+1.59)	79.03 (+1.59)
ResNet-18 (nasty)	77.42(-0.02)	64.49 (-6.68)	3.45 (-65.67)	74.81 (-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)

Data-Free KD

Table 6: Data-free KD from nasty teacher on CIFAR-10 and CIFAR-100

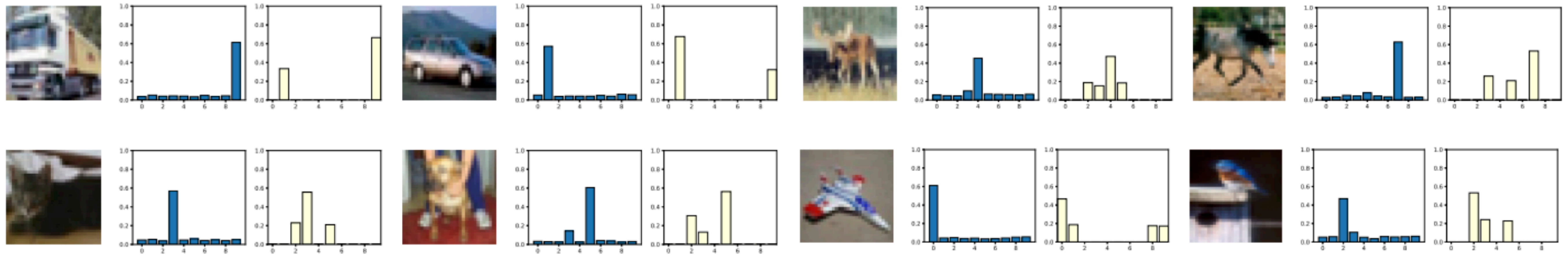
dataset	CIFAR-10		CIFAR-100	
Teacher Network	Teacher Accuracy	DAFL	Teacher Accuracy	DAFL
ResNet34 (normal)	95.42	92.49	76.97	71.06
ResNet34 (nasty)	94.54 (-0.88)	86.15 (-6.34)	76.12 (-0.79)	65.67 (-5.39)

(a) Normal Teacher



(b) Nasty Teacher

Analysis



- multi-peak logits may give a false sense of generalization and thus mislead the learning of students

Contact: {haoyum3,xhx}@uci.edu,{tianlong.chen, atlaswang}@utexas.edu, tkhu@tamu.edu, chenyu.you@yale.edu

Thanks!