Hardware Accelerated Learning

NEURAL INFORMATION PROCESSING SYSTEMS

# Analyzing the Confidentiality of Undistillable Teachers in Knowledge Distillation

*35th Neural Information Processing Systems, 2021*






**Souvik Kundu, Qirui Sun, Yao Fu, Massoud Pedram, Peter A. Beerel**

*Ming Hsieh Department of Electrical and Computer Engineering*

USC University of Southern California

# Machine Learning as a Service (MLAAS) is on the Rise

USC University of Southern California

Household robots

Autonomous driving

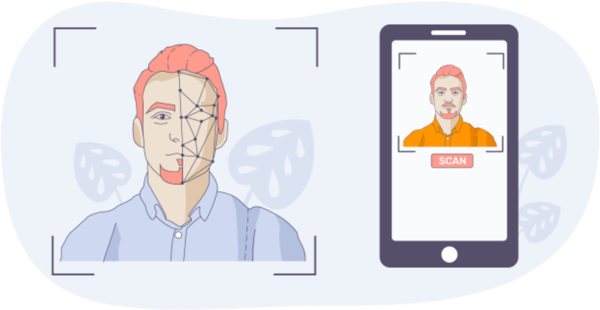



Image analysis

*Image courtesy: Google images*

- Various trained models are deployed at the edge to perform complex computer vision and natural language processing tasks

- Industries prefer the trained models to be released as commercial black-box APIs

# Model Performance Protection is Important

- Winning teams of AI competitions do **not** want their model performance to be replicated by opponents

- Industry releasing models as commercial black-box API do **not** want their model performance to be replicated by a potential competitor

- Commercial black-box ML APIs often require **large** human resource and training costs that the owner wants to be compensated for via MLAAS earnings


## Neural Networks

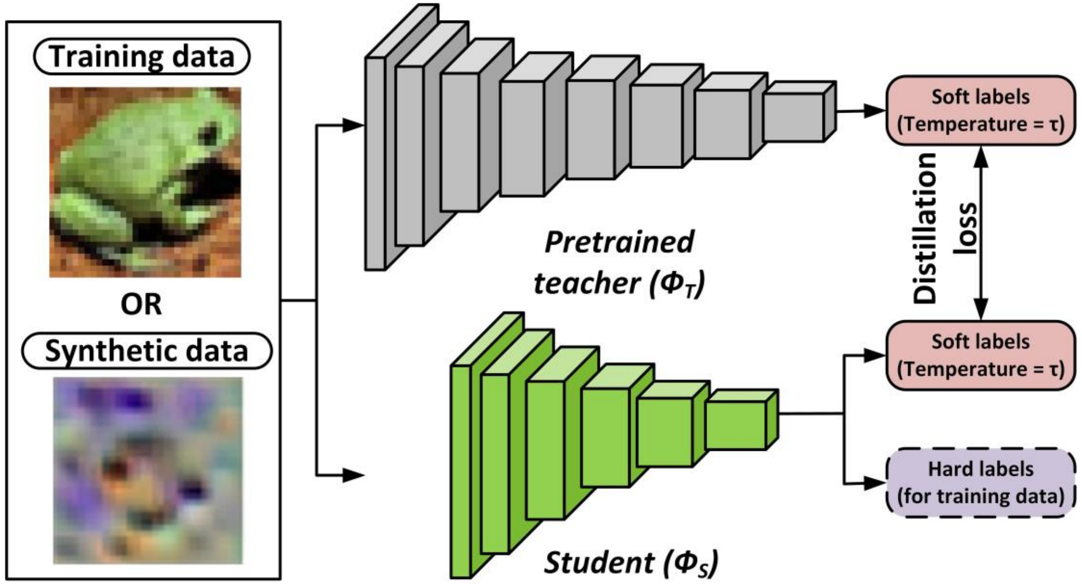
Apply cutting-edge research to train deep neural networks on problems ranging from perception to control. Our per-camera networks analyze raw images to perform semantic segmentation, object detection and monocular depth estimation. Our birds-eye-view networks take video from all cameras to output the road layout, static infrastructure and 3D objects directly in the top-down view. Our networks learn from the most complicated and diverse scenarios in the world, iteratively sourced from our fleet of nearly 1M vehicles in real time. A full build of Autopilot neural networks involves 48 networks that take 70,000 GPU hours to train 🔥. Together, they output 1,000 distinct tensors (predictions) at each timestep.

*Source: https://www.tesla.com/AI*

# Knowledge-Distillation (KD): A Potential Threat to MLAAS

USC University of Southern California

**Primary application: model compression**




**Concerning application: mimicking performance from black-box models**

- KD can transfer the "rich" knowledge of a compute-heavy teacher to a compute-efficient student model under both data-available[1] and data-free scenarios[2]

[1] Geoffrey Hinton et al., "Distilling the knowledge in a neural network", NeurIPS 2014 (workshop).
[2] Paul Micaelli and Amos Storkey, "Zero-shot knowledge transfer via adversarial belief matching", NeurIPS 2019.

# Undistillable Models[1]

USC University of Southern California

- A class of models that
  - Perform similar to standard teacher models to maintain their own performance
  - However, act as "**nasty**" teachers to any student model by not allowing it to mimic performance.
- Core idea
  - Inject **false** sense of generalization to the student[1]

Training loss of Undistillable models ($\boldsymbol{\Phi_T}$):

$$\boxed{\mathcal{L}_N = \underbrace{\mathcal{L}_{\mathcal{CE}}\big(\sigma(g_{\Phi_T}(\boldsymbol{x},\boldsymbol{y}))\big)}_{\text{Cross-entropy (CE) loss}} - \underbrace{\alpha_N * \tau_N^2 * \mathcal{L}_{\mathcal{KL}}\big(\sigma(g_{\Phi_T}(\boldsymbol{x},\boldsymbol{y}),\tau_N),\sigma(g_{\Phi_A}(\boldsymbol{x},\boldsymbol{y}),\tau_N)\big)}_{\text{Self-undermining loss}}}$$

*[1] Haoyu Ma et al., "Undistillable: Making a nasty teacher that cannot teach students", ICLR 2021 (spotlight).*

# A1: Analyzing Undistillability

USC University of Southern California

- A study of transferability of the impact of nasty teachers

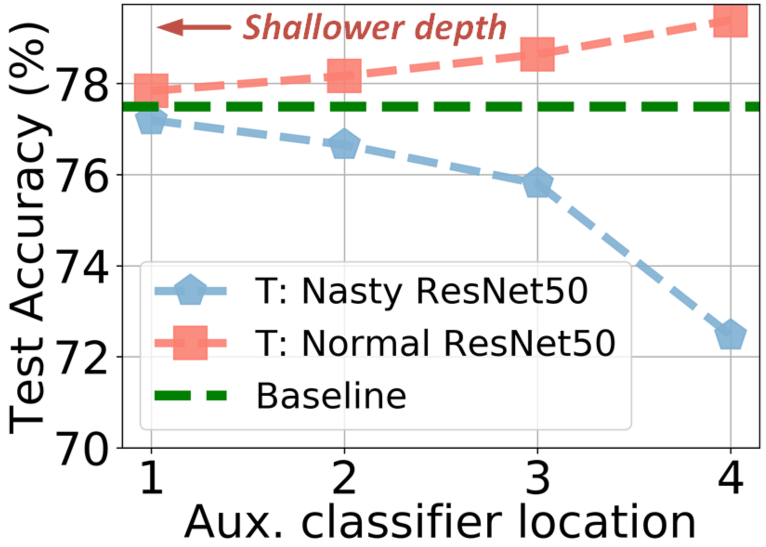
| Teacher | Teacher type | Teacher Acc % | Student Acc % | $\Delta_{base}$ |
|---|---|---|---|---|
| ResNet50 | Nasty | 76.57 | 72.47 | -5.08 |
| ResNet18 | Distilled | 72.47 | 70.99 | -6.56 |
| ResNet50 | Normal | 78.04 | 79.39 | +1.84 |
| ResNet18 | Distilled | 79.39 | 79.47 | +1.92 |

**The nastiness of a teacher transfers to its student**

# A2: Analyzing the Undistillability

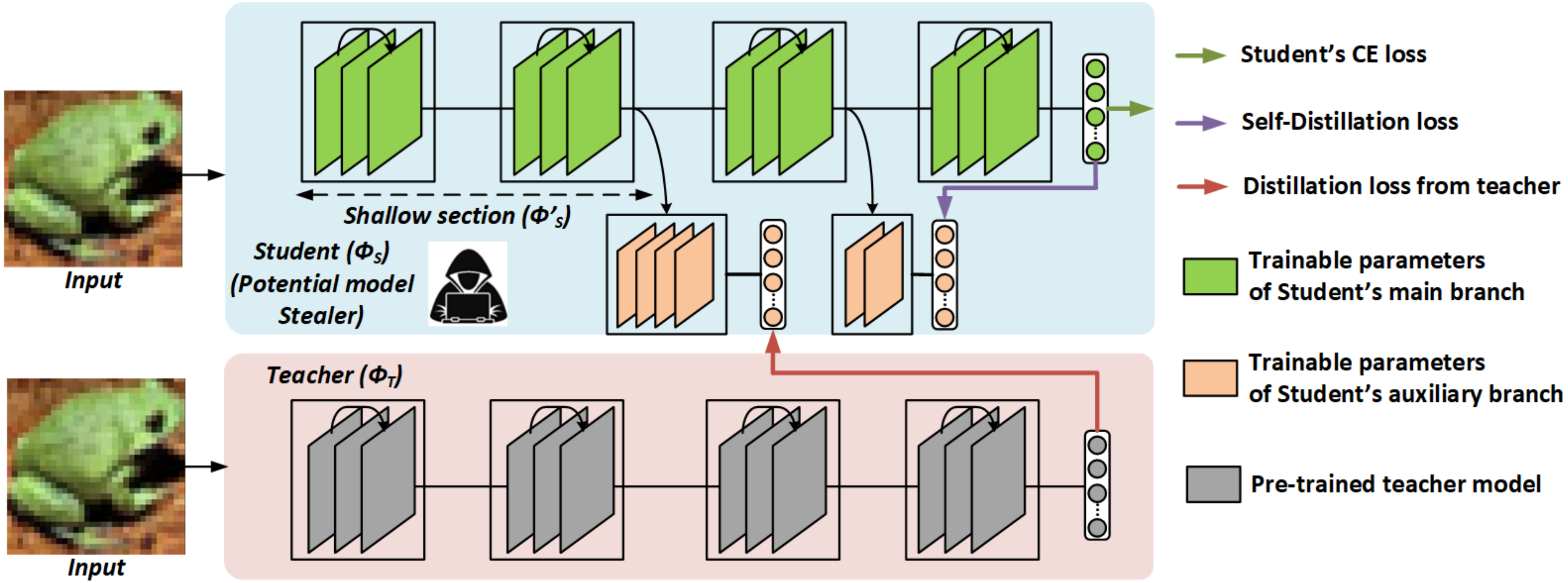
USC University of Southern California

- A study of applying KD at various depth of the student model




**Impact of a teacher reduces as we use KD at shallower depths of student**

# Our Proposal: Skeptical Student




- ➢ Transfer knowledge to shallow depth ($\Phi'_S$) of a student via aux. classifier (AC)
- ➢ Use self-distillation at AC in $\Phi_S$ - $\Phi'_S$ to boost performance of student $\Phi_S$

# Skeptical Students: Training Loss

USC University of Southern California

KL-divergence loss component:

$$\mathcal{L}_T = (1-\alpha) * \mathcal{L}_{\mathcal{CE}}\big(\sigma(g_{\Phi'_S}(\boldsymbol{x}, \boldsymbol{y}))\big) + \alpha * \tau^2 * \mathcal{L}_{\mathcal{KL}}\big(\sigma(g_{\Phi'_S}(\boldsymbol{x}, \boldsymbol{y}), \tau), \sigma(g_{\Phi_T}(\boldsymbol{x}, \boldsymbol{y}), \tau)\big)$$

Self-distillation loss component :

$$\mathcal{L}_{SD} = \sum_{j \in \mathcal{J}} \{(1-\beta) * \mathcal{L}_{\mathcal{CE}}\big(\sigma(g_{\Phi^j_S}(\boldsymbol{x}, \boldsymbol{y}))\big) + \beta * \mathcal{L}_{\mathcal{KL}}\big(\sigma(g_{\Phi^j_S}(\boldsymbol{x}, \boldsymbol{y}), \tau), \sigma(g_{\Phi_S}(\boldsymbol{x}, \boldsymbol{y}), \tau)\big)\}$$

CE loss component :

$$\mathcal{L}_{\mathcal{CE}}\big(\sigma(g_{\Phi_S}(\boldsymbol{x}, \boldsymbol{y}))\big)$$

Total loss (hybrid distillation):

$$\mathcal{L}_S = \gamma_1 \mathcal{L}_T + \gamma_2 \mathcal{L}_{SD} + \gamma_3 \mathcal{L}_{\mathcal{CE}}\big(\sigma(g_{\Phi_S}(\boldsymbol{x}, \boldsymbol{y}))\big)$$

# Skeptical Students: Distilled from Nasty Teachers

USC University of Southern California

| Dataset | $\Phi_T$ | $\Phi_T$ Acc. (%) | $\Phi_S$ | $\Phi_S$ Baseline Acc. (%) | Student Acc. (%) Normal ($acc_n$) | Skeptical ($acc_s$) | Skeptical-E ($acc_{se}$) | $\Delta_{acc}$ |
|---|---|---|---|---|---|---|---|---|
| CIFAR -10 | ResNet18 | 94.67 | ResNet18 | 95.15 | 94.13(±0.18) | **95.09**( ± 0.15) | 94.77(± 0.05) | +0.96 |
| | | | MobileNetV2 | 90.12 | 88.13(±0.13) | **90.37**( ± 0.25) | 90.21(± 0.18) | +2.24 |
| | ResNet50 | 94.28 | ResNet18 | 95.15 | 94.38(±0.18) | **95.16**( ± 0.01) | 95.02(± 0.01) | +0.78 |
| | | | ResNet50 | 94.9 | 94.21(±0.04) | **95.48**( ± 0.14) | 95.48(± 0.14) | +1.27 |
| | | | MobileNetV2 | 90.12 | 88.76(±0.14) | **91.02**( ± 0.09) | 90.88(± 0.23) | +2.26 |
| CIFAR -100 | ResNet18 | 77.55 | ResNet18 | 77.55 | 75.00(±0.14) | **77.33**( ± 0.21) | 76.38(± 0.1) | +2.33 |
| | | | MobileNetV2 | 69.24 | 7.13(±0.71) | **66.62**(± 0.30) | 64.26(± 0.64) | +59.49 |
| | ResNet50 | 76.57 | ResNet18 | 77.55 | 72.28(±0.27) | **77.25**( ± 0.25) | 75.48(±0.54) | +4.97 |
| | | | ResNet50 | 78.04 | 74.14(±0.85) | **78.65**( ± 0.29) | 77.61(±0.1) | +4.52 |
| | | | MobileNetV2 | 69.24 | 7.72(±1.57) | **66.38**( ± 0.50) | 62.93(±0.75) | +58.66 |
| Tiny-ImageNet | ResNet18 | 62.08 | ResNet18 | 63.07 | 53.60(±0.04) | **65.76**(±0.83) | 60.63(±0.07) | +12.16 |
| | | | MobileNetV2 | 57.01 | 4.81(± 0.19) | **54.74**(±0.84) | 54.27(±2.94) | +49.93 |

**Skeptical students achieve similar to teacher performance even when the teacher is Undistillable (or nasty).**

NEURAL INFORMATION PROCESSING SYSTEMS

# Skeptical Students: Distilled from Normal Teachers

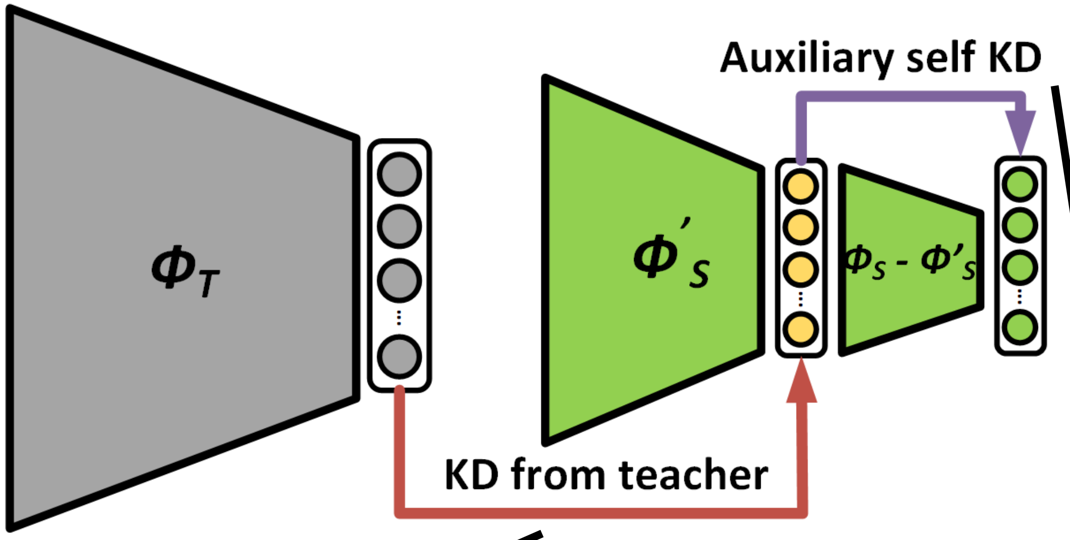
USC University of Southern California

| Dataset | $\Phi_T$ | $\Phi_T$ Acc. (%) | $\Phi_S$ | $\Phi_S$ Baseline Acc. (%) | Student Acc. (%) Normal ($acc_n$) | Skeptical ($acc_s$) | Skeptical-E ($acc_{s_e}$) | $\Delta_{acc}$ |
|---|---|---|---|---|---|---|---|---|
| CIFAR-10 | ResNet18 | 95.15 | ResNet18 | 95.15 | 95.38 (±0.10) | **95.45**(±0.10) | 95.42(±0.09) | +0.07 |
| | | | MobileNetV2 | 90.12 | 91.36(±0.17) | 91.81(±0.15) | **92.00**(±0.28) | +0.64 |
| | ResNet50 | 94.9 | ResNet18 | 95.15 | **95.43**(±0.11) | 95.31(±0.01) | 95.27(±0.04) | -0.12 |
| | | | ResNet50 | 94.9 | 95.15(±0.13) | 95.85(±0.05) | **96.09**(±0.01) | +0.94 |
| | | | MobileNetV2 | 90.12 | 91.71(±0.06) | 91.71(±0.18) | **91.95**(±0.16) | +0.24 |
| CIFAR-100 | ResNet18 | 77.55 | ResNet18 | 77.55 | 78.96(±0.12) | 78.79(±0.42) | **79.68**(±0.52) | +0.72 |
| | | | MobileNetV2 | 69.24 | 75.12(±0.08) | 71.63(±0.19) | **75.45**(±0.06) | +0.33 |
| | ResNet50 | 78.04 | ResNet18 | 77.55 | 79.21(±0.24) | 78.51(±0.44) | **79.86**(±0.01) | +0.65 |
| | | | ResNet50 | 78.04 | 79.56(±0.13) | 80.66(±0.52) | **81.96**(±0.52) | +2.4 |
| | | | MobileNetV2 | 69.24 | 75.28(±0.04) | 71.76(±0.16) | **76.32**(±0.34) | +1.04 |
| Tiny-ImageNet | ResNet18 | 63.07 | ResNet18 | 63.07 | 67.35(±0.18) | 66.49(±0.30) | **67.43**(±0.47) | +0.08 |
| | | | MobileNetV2 | 57.01 | 64.99(±0.51) | 59.37(±0.01) | **65.38**(±0.01) | +0.39 |

**Skeptical students achieve similar to normal students' performance upon distillation from a normal teacher.**

# Skeptical Students: Data-free Distillation






$$\mathcal{L}_{S_{DF}} = \mathcal{L}_{\mathcal{KL}}\big(\sigma(g_{\Phi'_S}(\boldsymbol{x},\boldsymbol{y}),\tau),\sigma(g_{\Phi_T}(\boldsymbol{x},\boldsymbol{y}),\tau)\big) + \mathcal{L}_{\mathcal{KL}}\big(\sigma(g_{\Phi_S}(\boldsymbol{x},\boldsymbol{y}),\tau),\sigma(g_{\Phi'_S}(\boldsymbol{x},\boldsymbol{y}),\tau)\big) + \boxed{\gamma_{at}\mathcal{L}_{\mathcal{AT}}}$$

Grey-box teacher: Attention transfer (AT) loss ✔

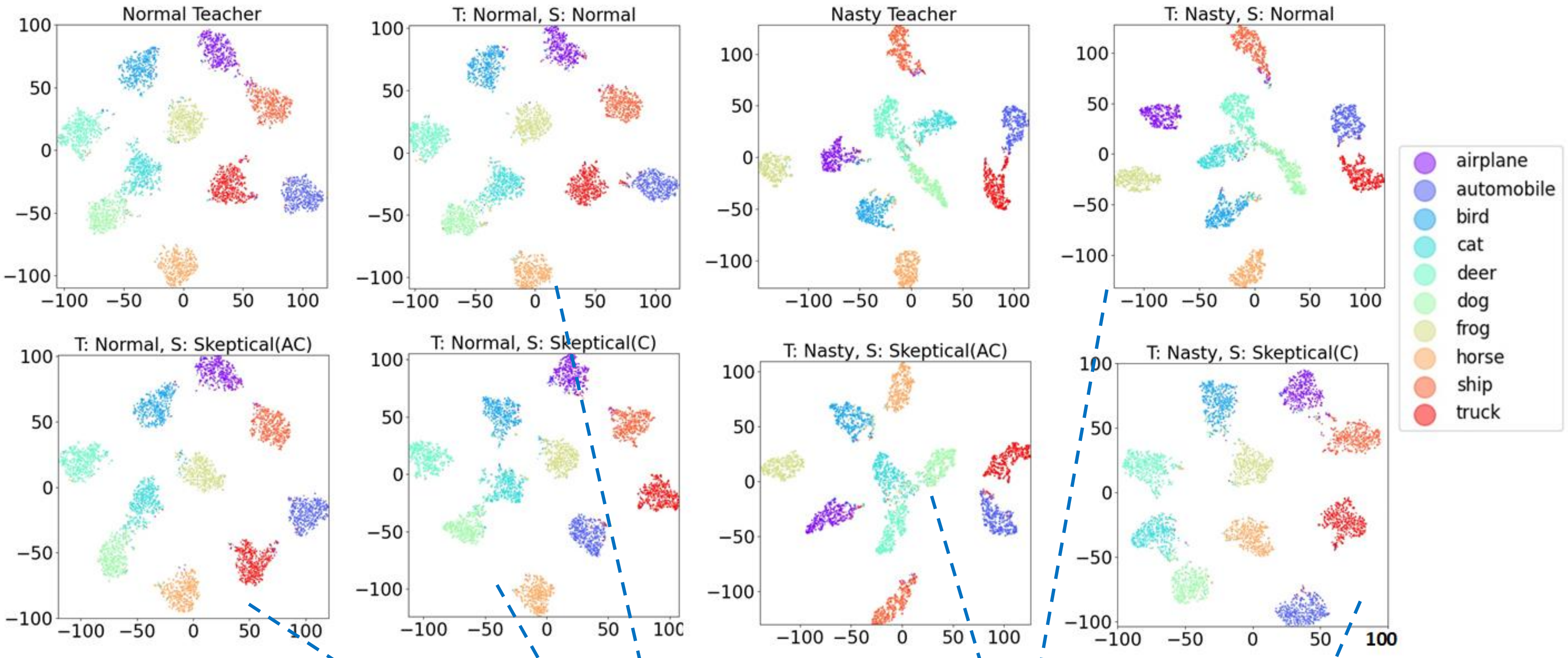
Black-box teacher: Attention transfer (AT) loss ✖

# Skeptical Students: Data-free Distillation Results

| Dataset | $\Phi_T$ | $\Phi_T$ type | $\Phi_T$ Acc. (%) | $\Phi_S$ | Student Acc. (%) | | $\Delta_{acc}$ |
|---|---|---|---|---|---|---|---|
| | | | | | Normal | Skeptical | |
| With AT loss (grey-box) | | | | | | | |
| CIFAR-10 | ResNet34 | Nasty | 94.81 | ResNet18 | 87.7(±1.20) | **91.76**(±0.30) | +4.06 |
| | | Normal | 95.3 | | 93.41(±0.21) | **93.52**(±0.06) | +0.11 |
| | ResNet50 | Nasty | 94.28 | | 80.34(±1.19) | **86.14**(±0.01) | +5.80 |
| | | Normal | 94.9 | | 90.54(±1.16) | **91.93**(±0.04) | +1.39 |
| Without AT loss (black-box) | | | | | | | |
| CIFAR-10 | ResNet50 | Nasty | 94.28 | ResNet18 | 20.95(±0.21) | **79.93**(±1.58) | **+58.98** |
| | | Normal | 94.9 | | 22.08(±0.56) | **80.71**(±1.21) | +58.63 |

Skeptical students achieve **significantly superior** performance compared to normal counter parts.

# Skeptical Students: Analysis of Results



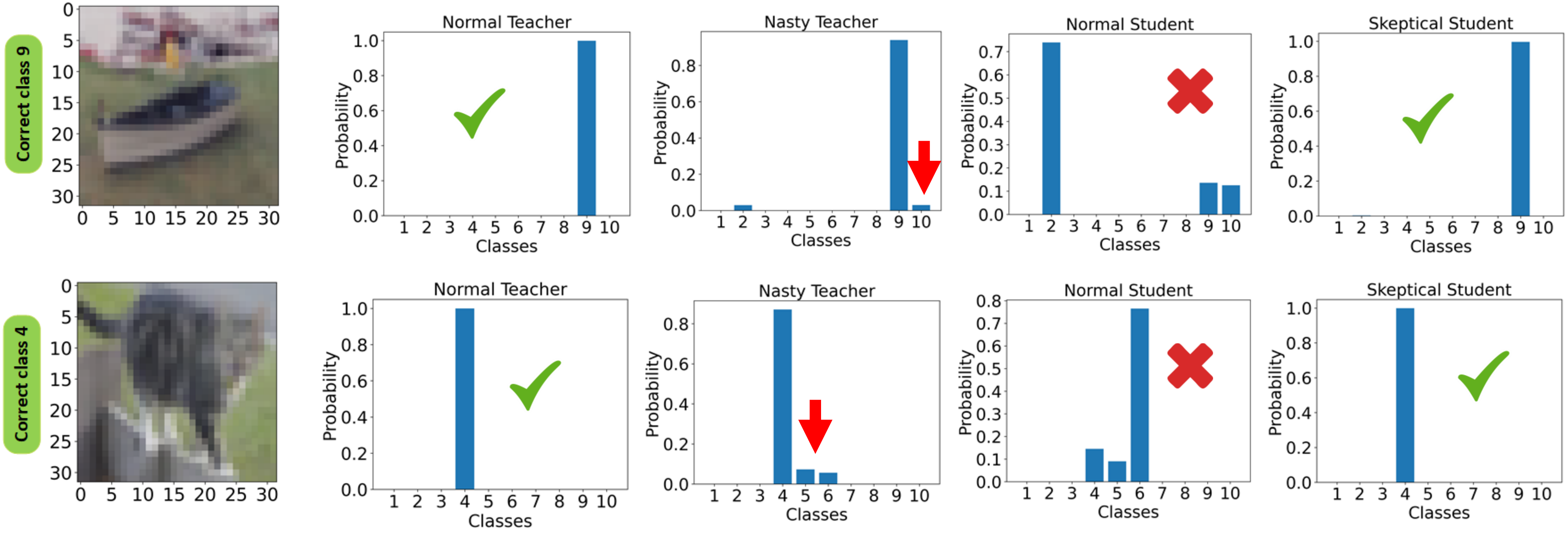

**Adverse mixing of class clusters**

**No visible adverse mixing of class clusters**

*Evaluations done on CIFAR-10 dataset with ResNet50 as teacher and ResNet18 as student model.*

# Skeptical Students: Analysis of Results

USC University of Southern California




Negligible logit values of incorrect classes

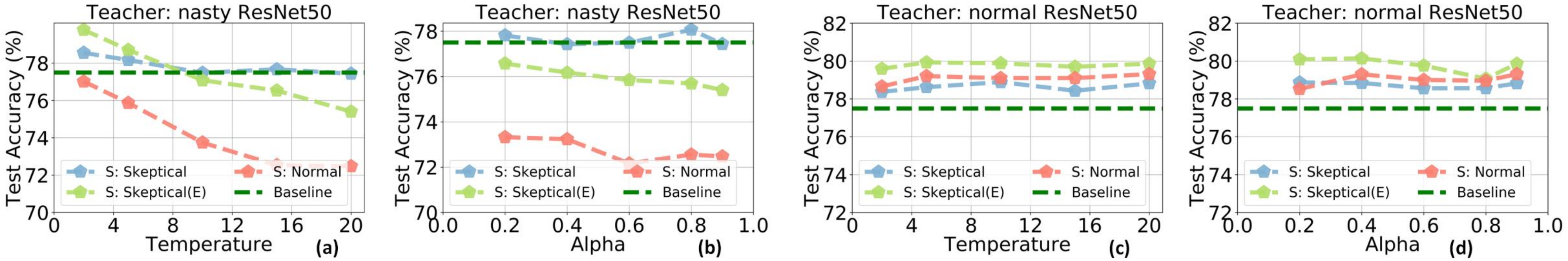
Non-negligible logit values of incorrect classes

Incorrectly classified class

*Evaluations done on CIFAR-10 dataset with ResNet50 as teacher and ResNet18 as student model.*

NEURAL INFORMATION PROCESSING SYSTEMS

# Skeptical Students: Ablation with Hyperparameters

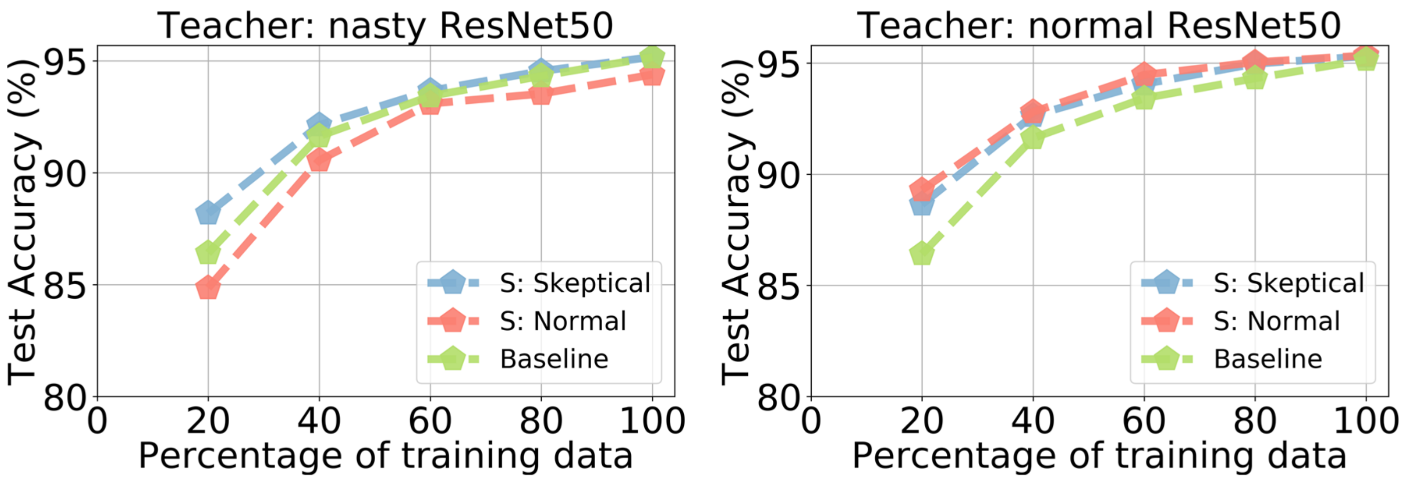



**Skeptical students consistently outperform normal counter parts on different loss strength and temperature value choices[1].**

[1] *Evaluation done on CIFAR-100 dataset to ResNet18 student model.*

# Skeptical Students: Ablation with Limited Data-availability

USC University of Southern California




**Skeptical students consistently outperform normal counter parts on various limited data availability scenarios[1].**

*[1] Evaluation done on CIFAR-10 dataset to ResNet18 student model.*

# Skeptical Students: Transferability of Nastiness

USC University of Southern California

| Teacher | Teacher type | Teacher Acc % | Student Acc % | $\Delta_{base}$ |
|---|---|---|---|---|
| ResNet50 | Nasty | 76.57 | 77.43 | -0.12 |
| ResNet18 | Nasty-distilled | 77.43 | 79.22 | +1.67 |
| ResNet50 | Normal | 78.04 | 78.90 | +1.35 |
| ResNet18 | Normal-distilled | 78.90 | 79.92 | +2.37 |

**The nastiness of a teacher does not get transferred to the skeptical student**

# Summary

USC University of Southern California

- Skeptical students can successfully **distill from even a nasty teacher** outperforming normal student counterparts

- Skeptical students can yield **better performance** on both data-available and data-free scenarios

- The success of skeptical students in mimicking model performance **poses a fundamental question** on protecting model IP in a distillation framework.

**Acknowledgment: This work was supported in parts by NSF including grant number 1763747.**

# Thank You!