南京航空航天大學

Nanjing University of Aeronautics and Astronautics

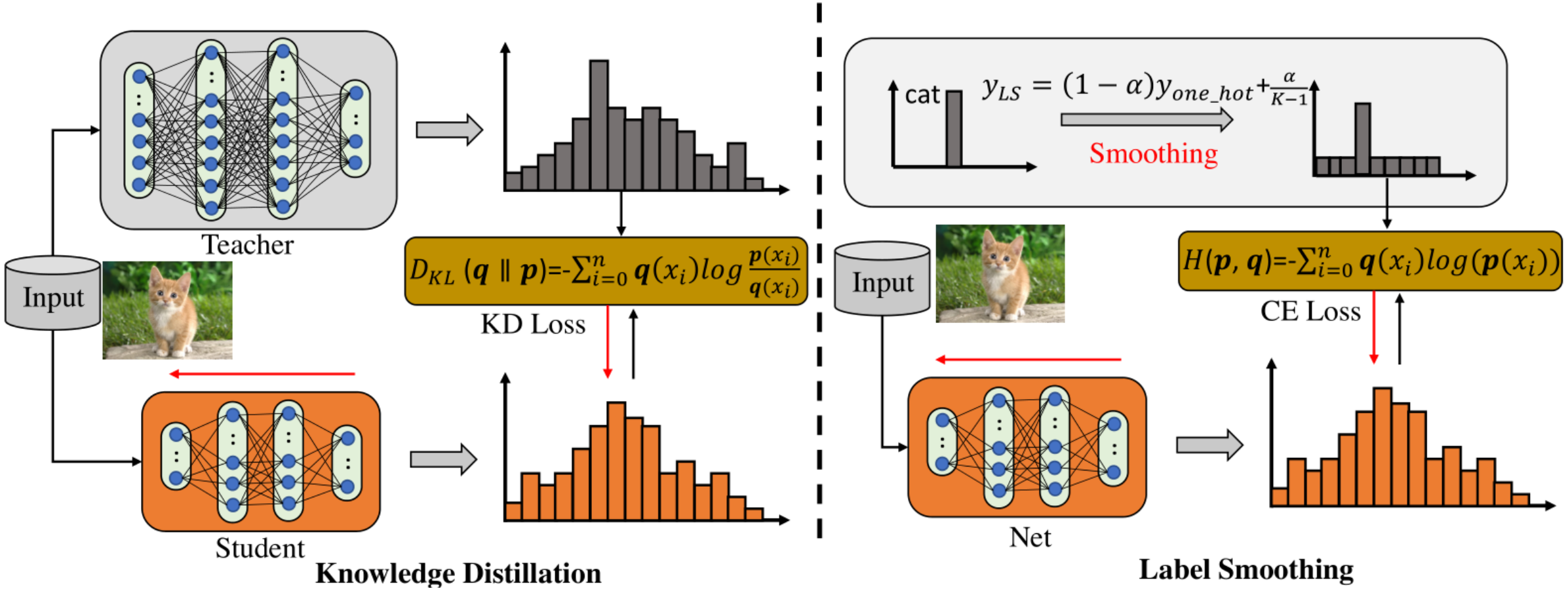
ParNeC 模式识别与神经计算研究组

PAttern Recognition and NEural Computing

Two Articles about

Label Smoothing and Knowledge Distillation

Knowledge Distillation & Label Smoothing



$$\lambda \mathcal{H}(\boldsymbol{p}^{\mathcal{S}_{\mathbf{w}}}, \boldsymbol{y}) + (1-\lambda)\mathcal{H}(\boldsymbol{p}^{\mathcal{S}_{\mathbf{w}}}/\mathcal{T}, \boldsymbol{p}^{\mathcal{T}_{\mathbf{w}}}/\mathcal{T})$$

When Does Label Smoothing Help?

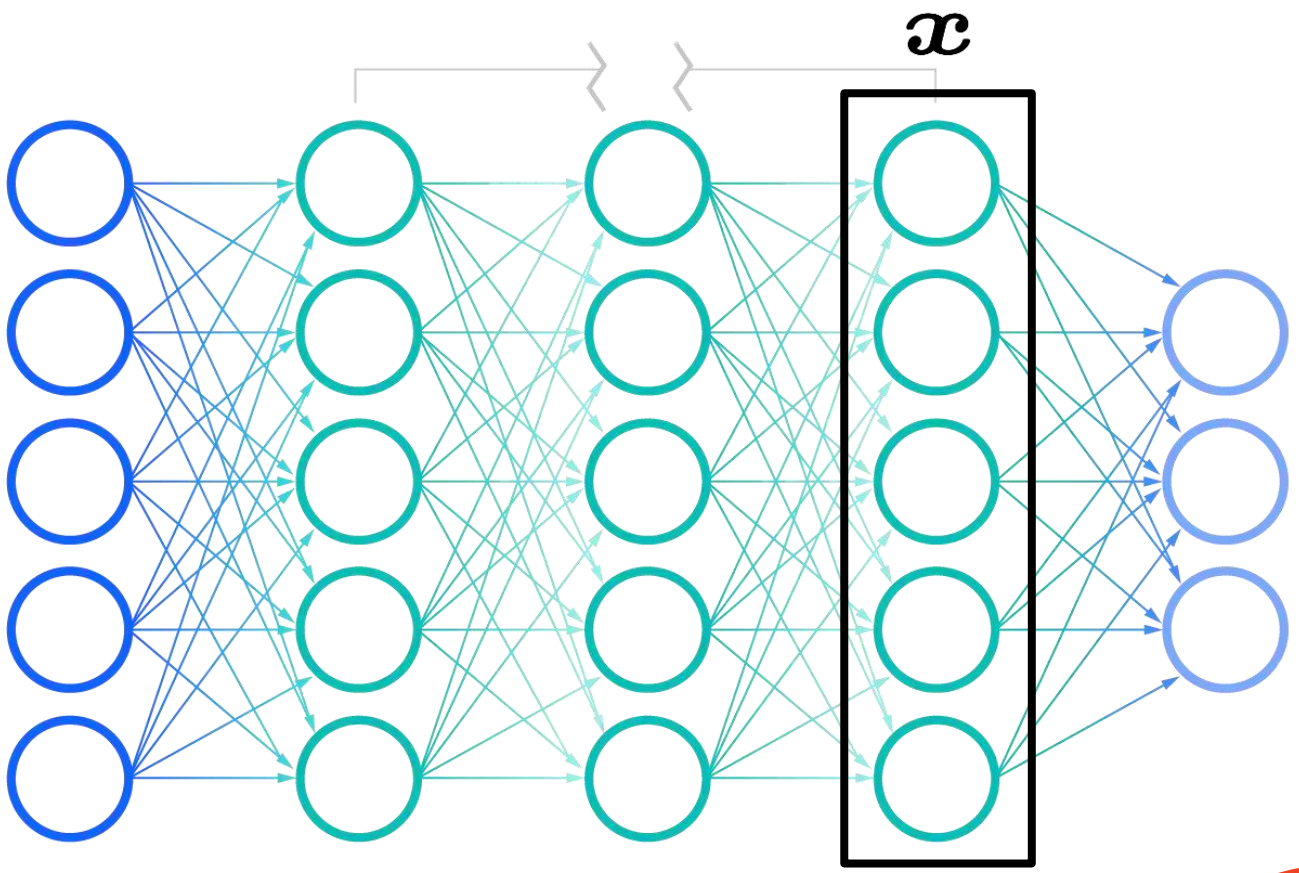
Rafael Müller,* Simon Kornblith, Geoffrey Hinton
Google Brain
Toronto
rafaelmuller@google.com

NIPS 2019

Survey of Literature Label Smoothing Results

Data set	Architecture	Metric	Value w/o LS	Value w/ LS
ImageNet	Inception-v2 [6]	Top-1 Error	23.1	**22.8**
		Top-5 Error	6.3	**6.1**
EN-DE	Transformer [11]	BLEU	25.3	**25.8**
		Perplexity	**4.67**	4.92
WSJ	BiLSTM+Att.[10]	WER	8.9	7.0/**6.7**

Preliminaries



$$p_k = \frac{e^{\boldsymbol{x}^T \boldsymbol{w}_k}}{\sum_{l=1}^{L} e^{\boldsymbol{x}^T \boldsymbol{w}_l}}$$

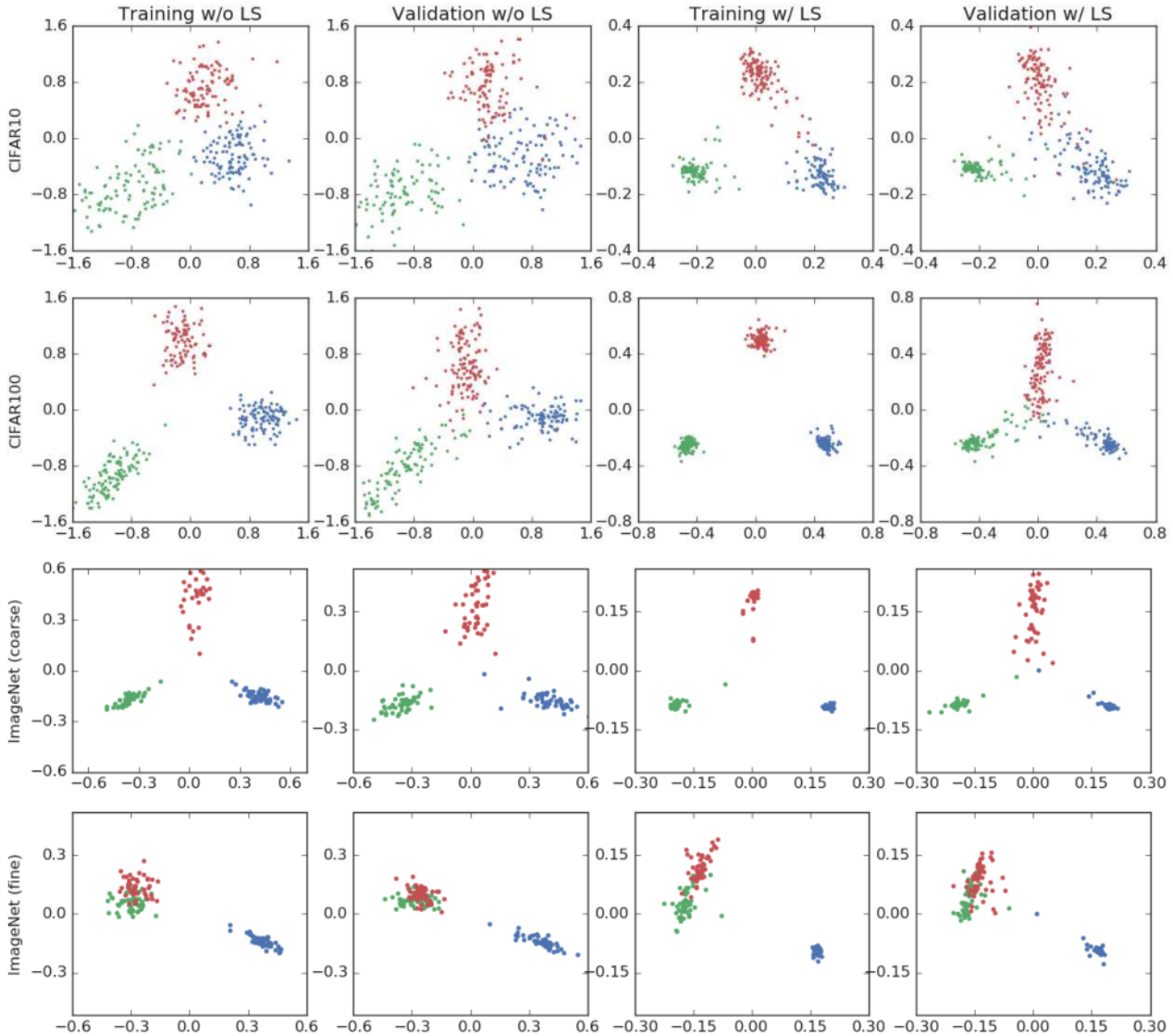
$$H(\boldsymbol{y}, \boldsymbol{p}) = \sum_{k=1}^{K} -y_k \log(p_k)$$

$$\|\boldsymbol{x} - \boldsymbol{\omega_k}\|^2 = \boldsymbol{x^T x} - 2\boldsymbol{x^T \omega_k} + \boldsymbol{\omega_k^T \omega_k}$$

Factored out

Usually constant across classes

Penultimate Layer Representations

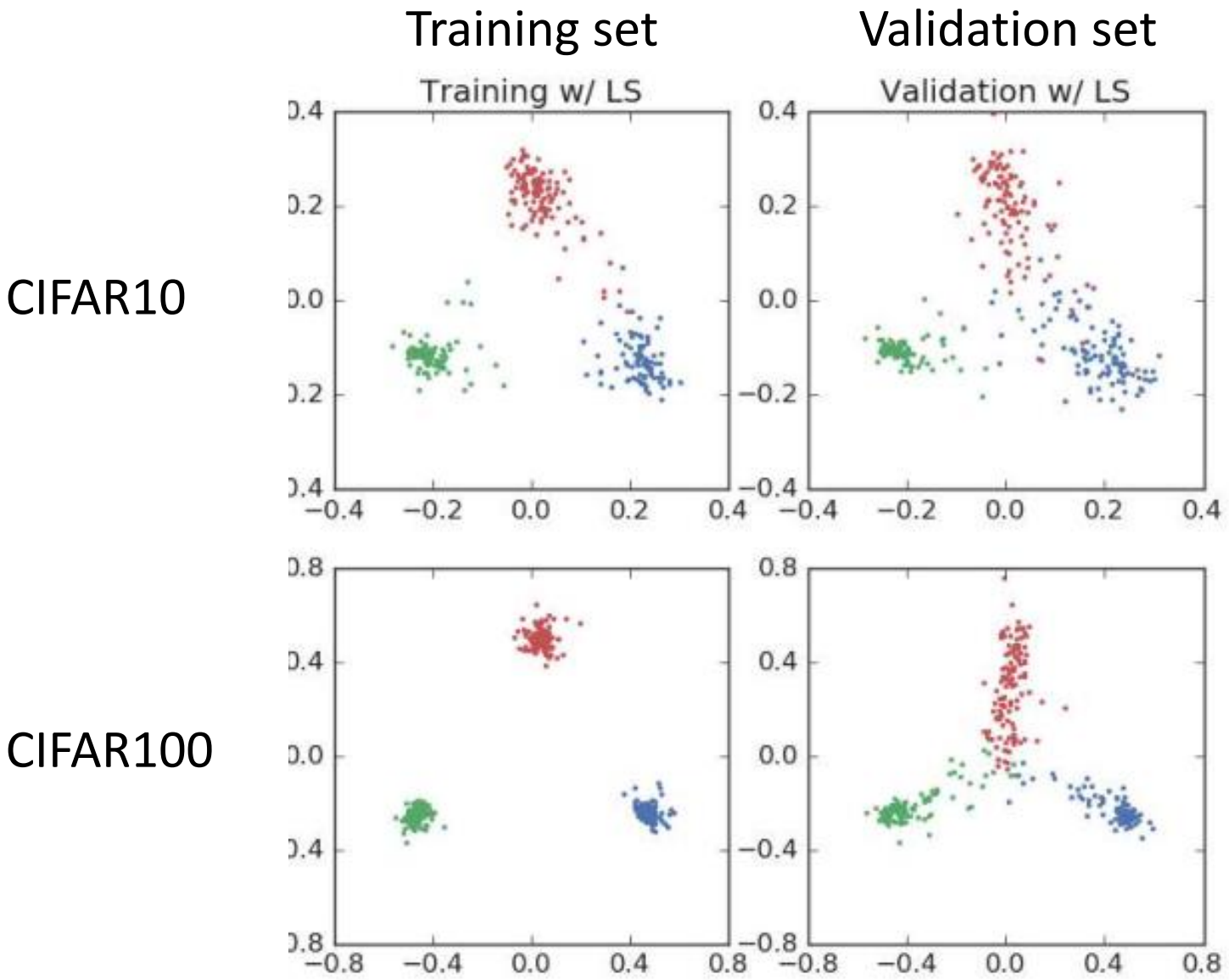


3 semantically different classes

2 similar classes + 1 class

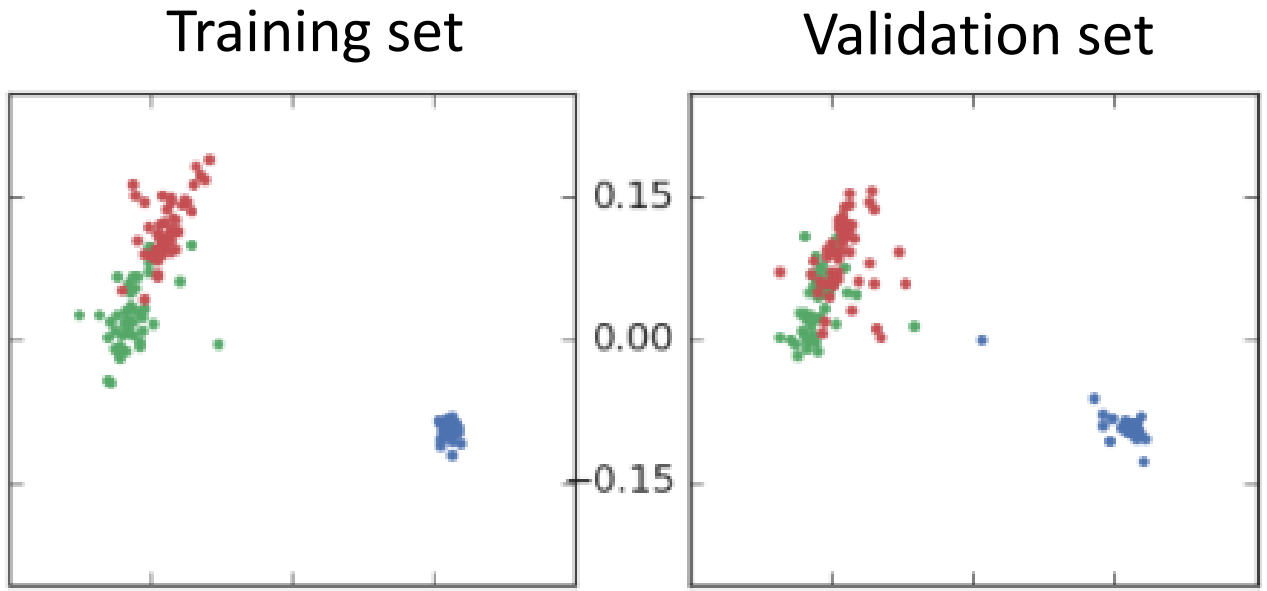
Figure 1: Visualization of penultimate layer's activations of: AlexNet/CIFAR-10 (first row), CIFAR-100/ResNet-56 (second row) and ImageNet/Inception-v4 with three semantically different classes (third row) and two semantically similar classes plus a third one (fourth row).

Penultimate Layer Representations



Penultimate Layer Representations

ImageNet



2 similar classes + 1 class

Penultimate Layer Representations

Data set	Architecture	Accuracy ($\alpha = 0.0$)	Accuracy ($\alpha = 0.1$)
CIFAR-10	AlexNet	86.8 ± 0.2	86.7 ± 0.3
CIFAR-100	ResNet-56	72.1 ± 0.3	72.7 ± 0.3
ImageNet	Inception-v4	80.9	80.9

Implicit Model Calibration

	$0 \le \hat{p}(\mathbf{x}) < 0.2$	$0.2 \le \hat{p}(\mathbf{x}) < 0.4$	$0.4 \le \hat{p}(\mathbf{x}) < 0.6$	$0.6 \le \hat{p}(\mathbf{x}) < 0.8$	$0.8 \le \hat{p}(\mathbf{x}) < 1$
Avg conf:			0.55	0.74	0.86
— Accuracy:			0.50	0.67	0.71
Gap:			2 x 0.05	6 x 0.17	7 x 0.15

$$\frac{2 \times 0.05 + 6 \times 0.17 + 7 \times 0.15}{15}$$

Expected Calibrated Error (ECE) = 0.11

Implicit Model Calibration

Data set	Architecture	Baseline ECE (T=1.0, $\alpha = 0.0$)	Temp. scaling ECE / T ($\alpha = 0.0$)	Label smoothing ECE / α (T=1.0)
CIFAR-100	ResNet-56	0.150	0.021 / 1.9	0.024 / 0.05
ImageNet	Inception-v4	0.071	0.022 / 1.4	0.035 / 0.1
EN-DE	Transformer	0.056	0.018 / 1.13	0.019 / 0.1

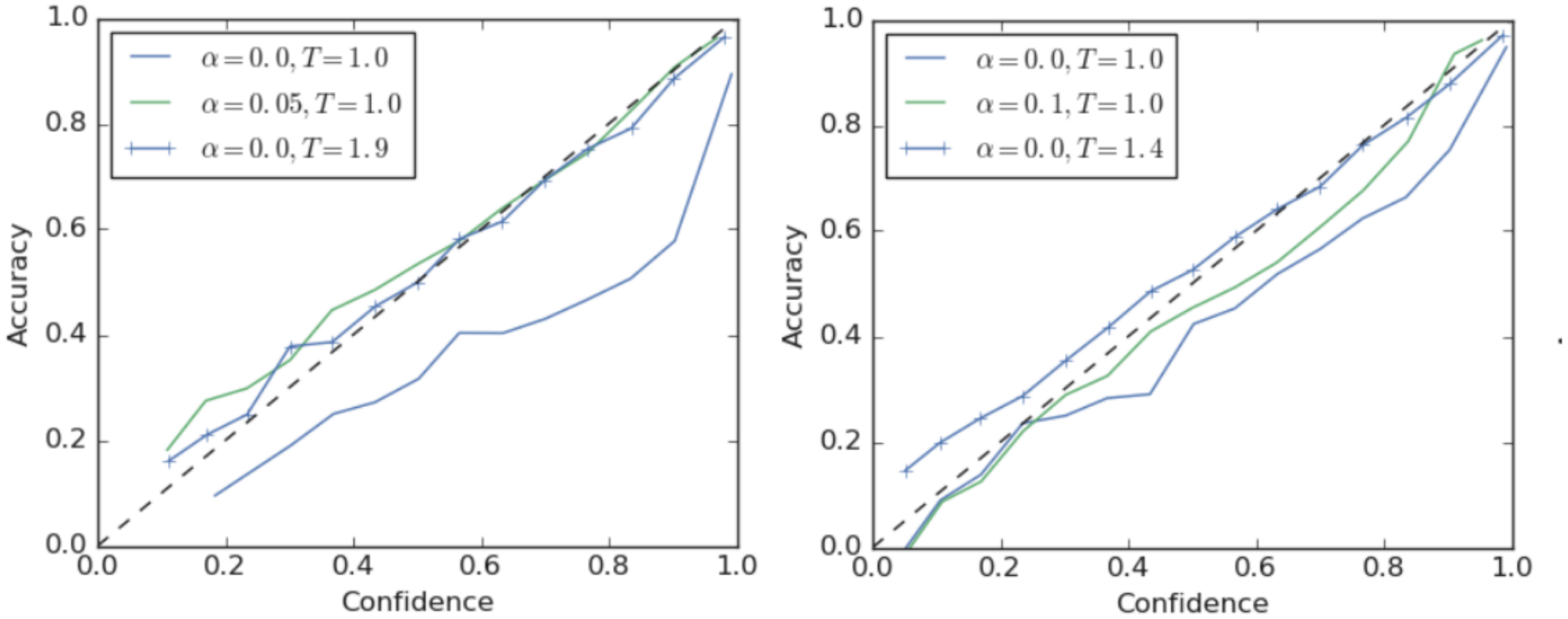


Figure 2: Reliability diagram of ResNet-56/CIFAR-100 (left) and Inception-v4/ImageNet (right).

Implicit Model Calibration

The setting of machine translation experiment is interesting for two reasons:

1. Label Smoothing → BLEU↑ but Perplexity↑

2. CV: only care about accuracy
   NLP: the network's soft outputs are inputs to a second algorithm(beam-search) which is affected by calibration

Implicit Model Calibration

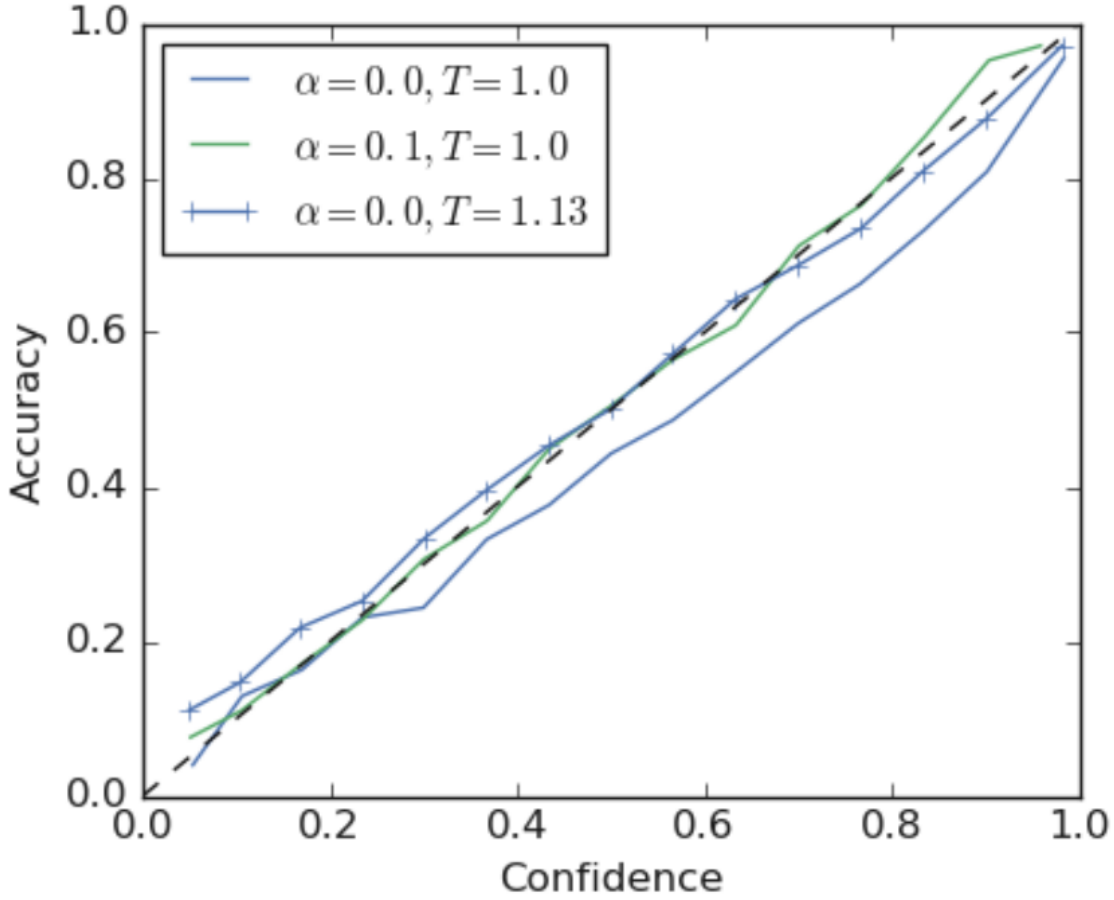


Figure 3: Reliability diagram of Transformer trained on EN-DE dataset.

Implicit Model Calibration

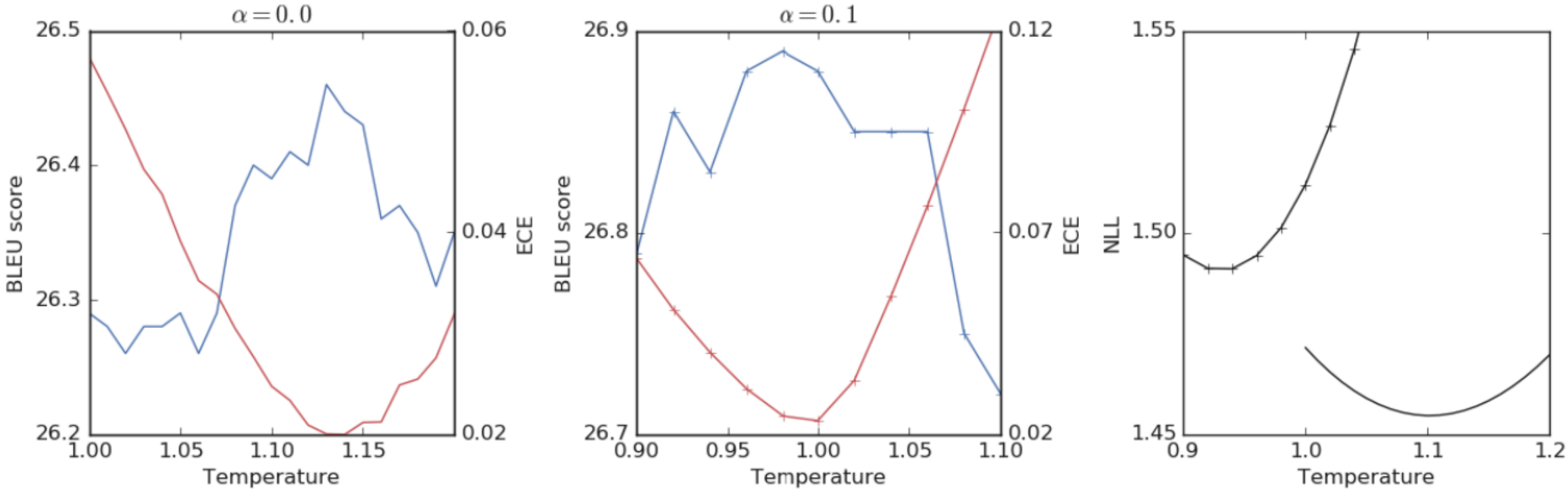


Figure 4: Effect of calibration of Transformer upon BLEU score (blue lines) and NLL (red lines). Curves without markers reflect networks trained without label smoothing while curves with markers represent networks with label smoothing.

Knowledge Distillation

Experiments on MNIST

T/S	Settings	Test error
Teacher	non-convolutional + hard targets + **dropout**	0.67%
Student	/	**0.74%**
Teacher	non-convolutional + hard targets + **label smoothing**	**0.59%**
Student	/	0.91%

$$(1-\beta)H(\boldsymbol{y},\boldsymbol{p})+\beta H(\boldsymbol{p}^t(T),\boldsymbol{p}(T))$$

Knowledge Distillation

Focus on four results:

1. the teacher's accuracy as a function of the label smoothing factor

2. the student's baseline accuracy as a function of the label smoothing factor without distillation

3. the student's accuracy after distillation with temperature scaling to control the smoothness of the teacher's provided targets (teacher trained with hard targets)

4. the student's accuracy after distillation with fixed temperature (T = 1.0 and teacher trained with label smoothing to control the smoothness of the teacher's provided targets)

Knowledge Distillation

$$\gamma = \mathbb{E}\left[\sum_{k=1}^{K}(1-y_k)p_k^t(T)K/(K-1)\right]$$

the mass allocated by the teacher to incorrect examples over training set

$$y_k^{LS} = y_k(1-\alpha) + \alpha/K$$

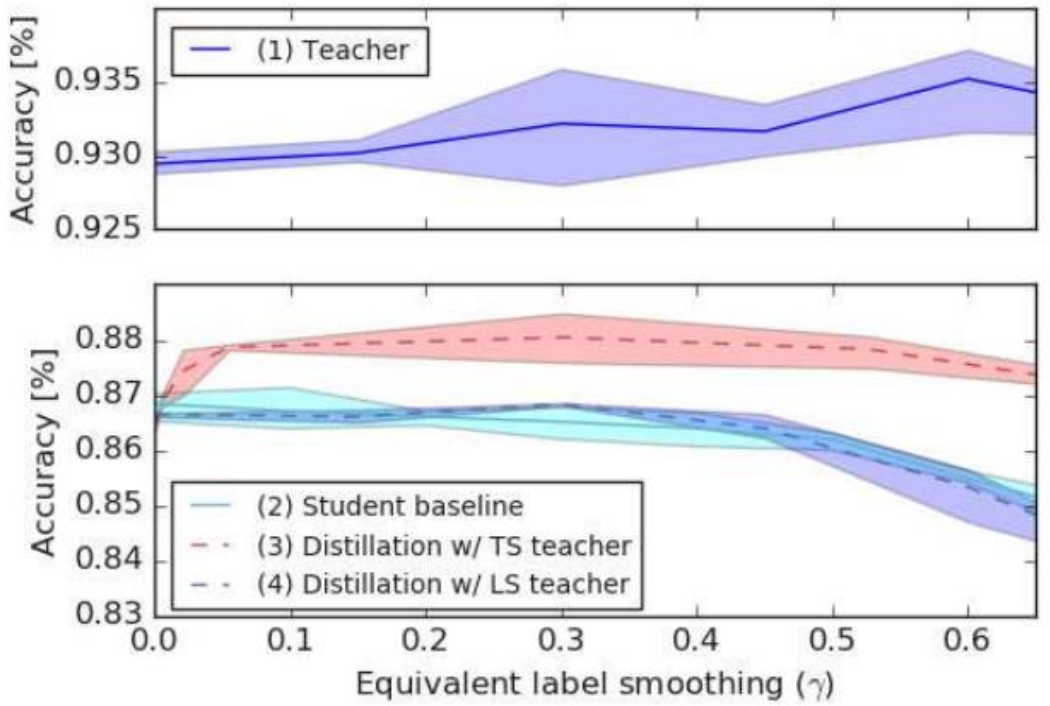


Figure 5: Performance of distillation from ResNet-56 to AlexNet on CIFAR-10.

Knowledge Distillation

$$y = f(d(\boldsymbol{z}_x))$$

$$I(X;Y) = E_{X,Y}[\log(p(y|x)) - \log(\textstyle\sum_x p(y|x))]$$

$$\hat{I}(X;Y) = \sum_{x=1}^{N}\big[-(f(d(\boldsymbol{z}_x)) - \mu_x)^2/(2\sigma^2)$$

$$- \log(\sum_{x=1}^{N} e^{-(f(d(\boldsymbol{z}_x))-\mu_x)^2/(2\sigma^2)})\big]$$

where

$$\mu_x = \textstyle\sum_{l=1}^{L} f(d(\boldsymbol{z}_x))/L,\ \sigma^2 = \sum_{x=1}^{N}(f(d(\boldsymbol{z}_x)) - \mu_x)^2/N$$

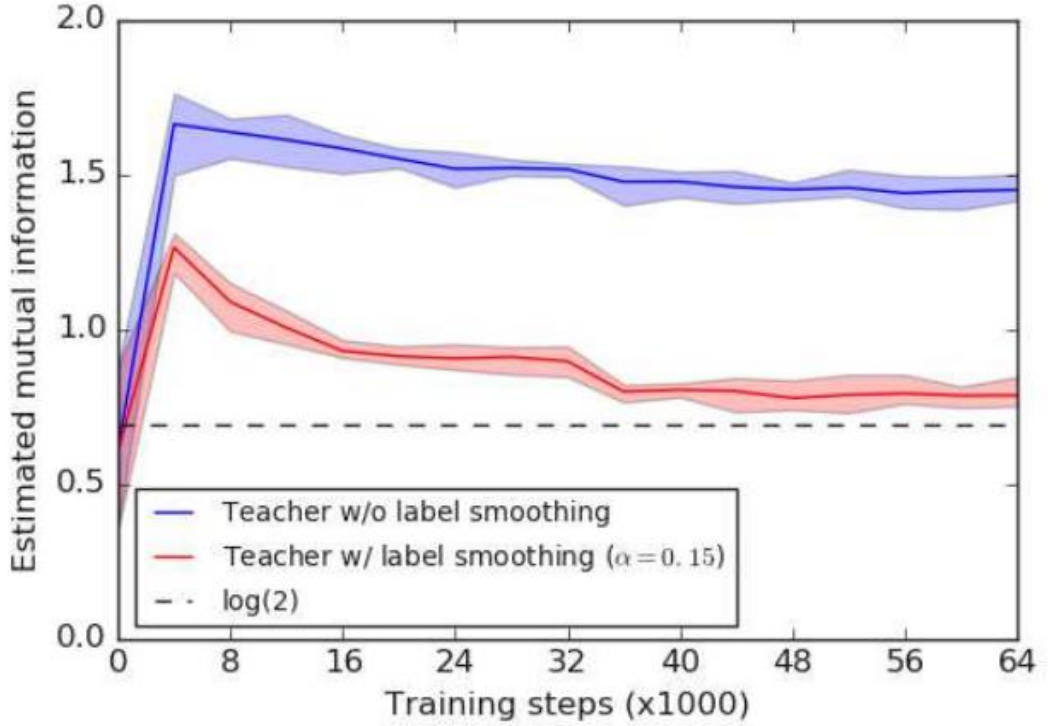


Figure 6: Estimated mutual information evolution during teacher training.

IS LABEL SMOOTHING TRULY INCOMPATIBLE WITH KNOWLEDGE DISTILLATION: AN EMPIRICAL STUDY

Zhiqiang Shen
CMU

Zechun Liu
CMU & HKUST

Dejia Xu
Peking University

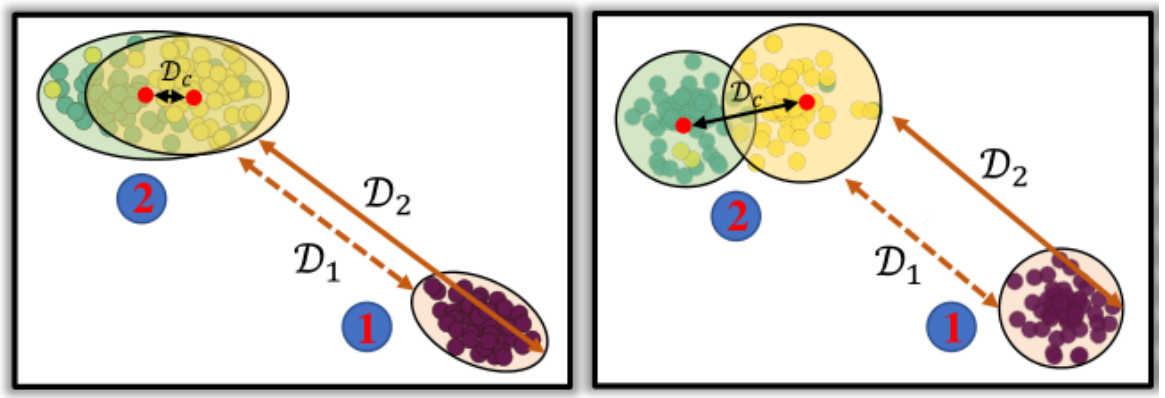
Zitian Chen
UMass Amherst

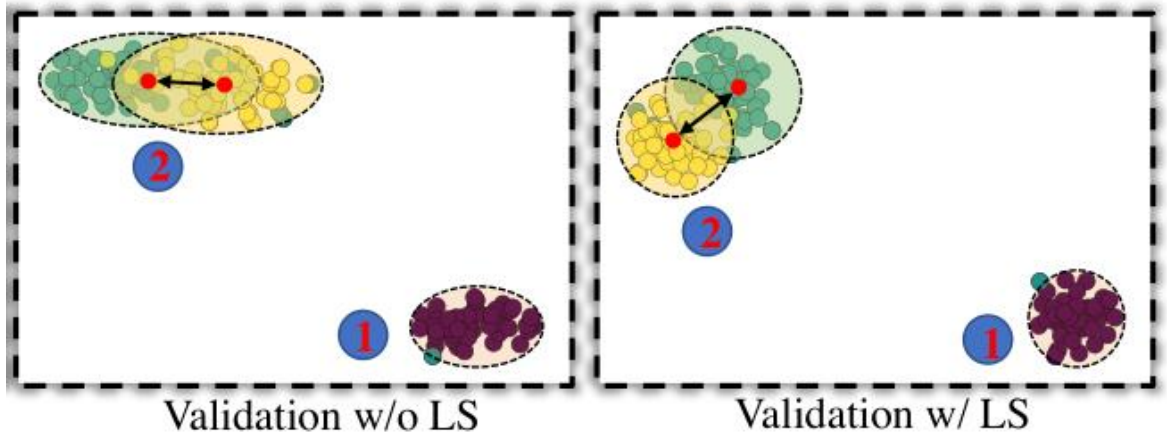
Kwang-Ting Cheng
HKUST

Marios Savvides
CMU

ICLR 2021

The Most Important Picture in this Article





Picture in the First Article

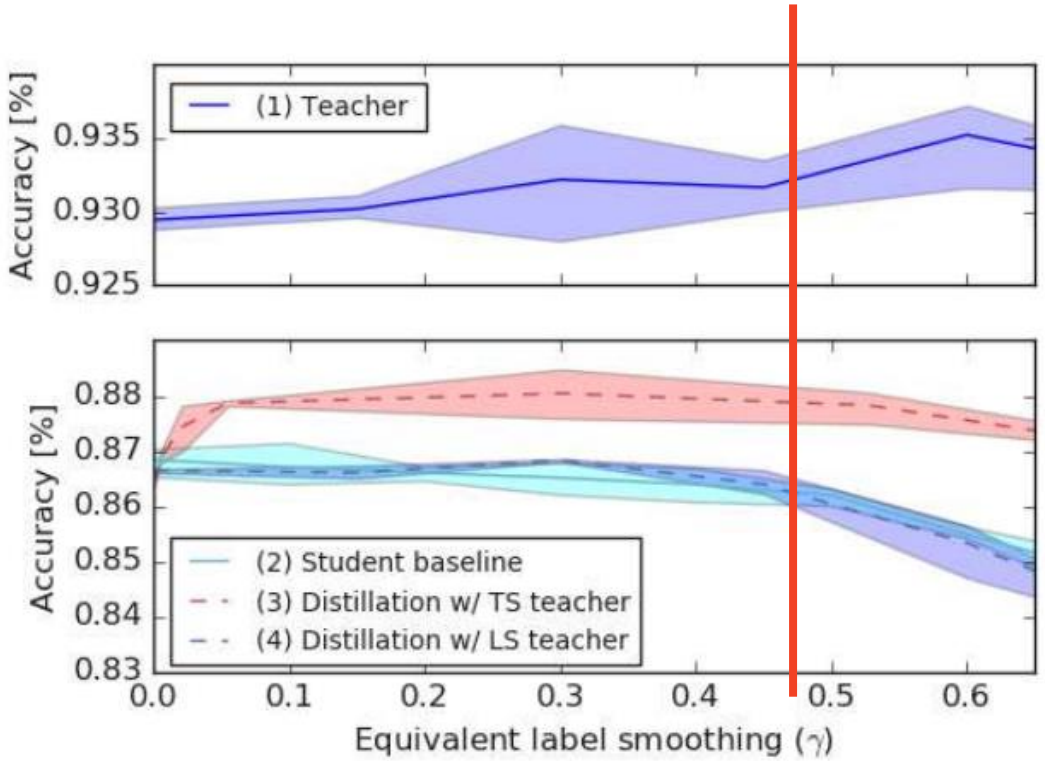
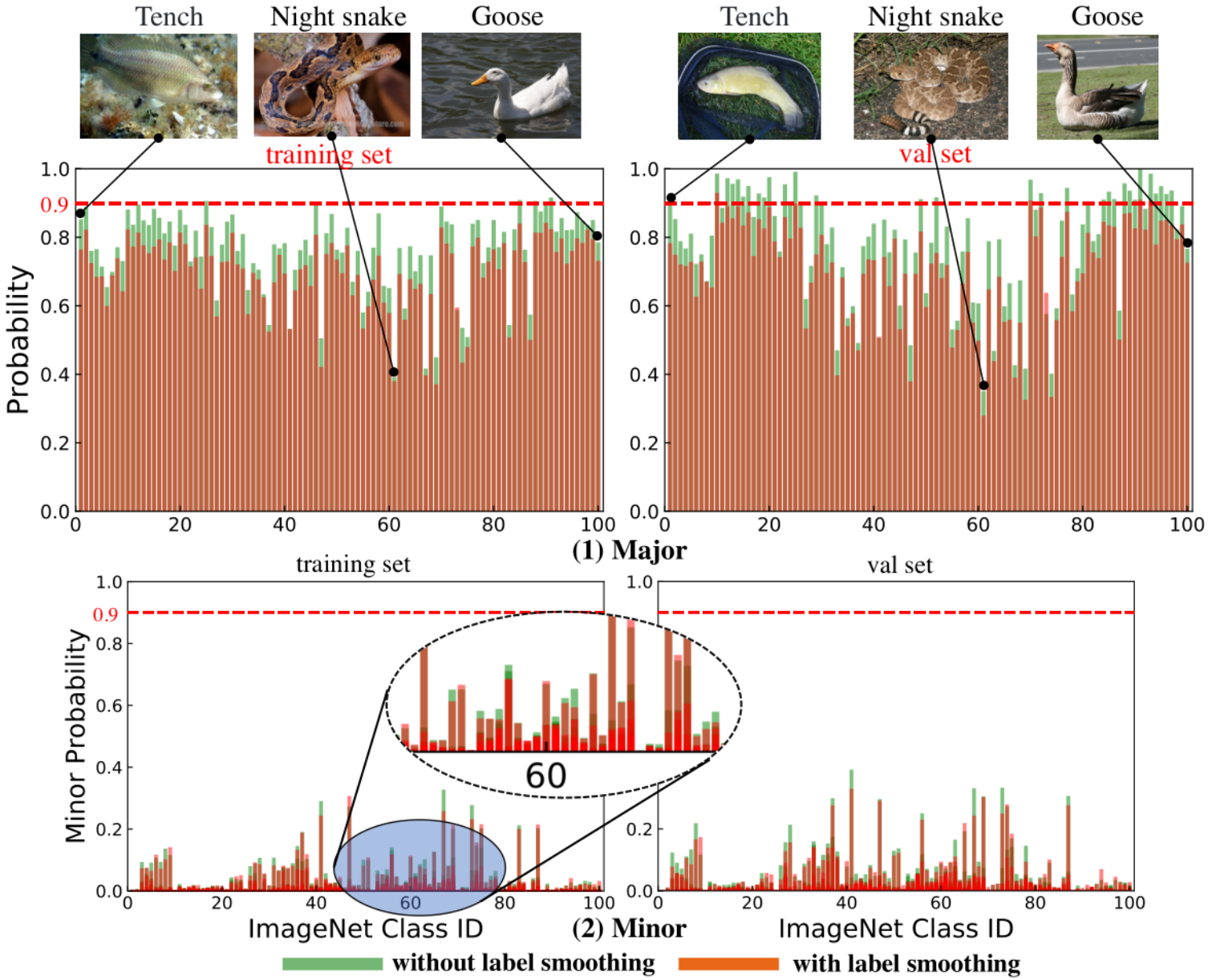


Figure 5: Performance of distillation from ResNet-56 to AlexNet on CIFAR-10.

$$\gamma = \mathbb{E}\left[\sum_{k=1}^{K}(1-y_k)p_k^t(T)K/(K-1)\right]$$

The “Erase Information” Effect by Label Smoothing



The “Erase Information” Effect by Label Smoothing

$$\mathcal{S}_{\text{Stability}} = 1 - \frac{1}{K}\sum_{c=1}^{K}\left(\frac{1}{\boldsymbol{n}_c}\sum_{i=1}^{\boldsymbol{n}_c}||\boldsymbol{p}^{\mathcal{T}_{\mathbf{w}}}_{\{i,c\}} - \overline{\boldsymbol{p}}^{\mathcal{T}_{\mathbf{w}}}_{\{i,c\}}||^2\right)$$

Some interesting facts(the authors do not give specific experimental data):

1) It can measure the degree of erased information quantitatively and further help discover more interesting phenomena, e.g., they observe that data augmentation method like CutMix (Yun et al., 2019) together with longer training erases the relative information on logits dramatically and can further be reinforced by label smoothing.

2) We found that the proposed metric is highly aligned with model accuracy, thus such metric can be used as a complement for accuracy to evaluate the quality of teacher’s supervision for knowledge distillation.

Empirical Studies

~~Netowrks~~ Networks	Acc. (%) w/o LS	(1-$\mathcal{S}_{\text{Stability}}$) w/o LS	Acc. (%) w/ LS	(1-$\mathcal{S}_{\text{Stability}}$) w/ LS
ResNet-18 (He et al., 2016)	69.758/89.078	0.3359	**69.774/89.122**	**0.3358 (-0.0001)**
ResNet-50 (He et al., 2016)	75.888/92.642	0.3217	**76.130/92.936**	**0.3106 (-0.0111)**
ResNet-101 (He et al., 2016)	77.374/93.546	0.3185	**77.726/93.830**	**0.3070 (-0.0115)**
MobileNet v2 (Sandler et al., 2018)	71.878/90.286	0.3341	–	–
DenseNet121 (Huang et al., 2017)	74.434/91.972	0.3243	–	–
ResNeXt50 32×4d (Xie et al., 2017)	77.618/93.698	0.3229	**77.774/93.642**	**0.3182 (-0.0047)**
Wide ResNet50 (Zagoruyko & Komodakis, 2016)	**78.468/94.086**	0.3201	77.808/93.682	**0.3155 (-0.0046)**
ResNeXt101 32×8d (Xie et al., 2017)	79.312/94.526	0.3177	**79.698/94.768**	**0.3116 (-0.0061)**
ResNet50+Long	76.526/93.070	0.3222	**77.106/93.340**	**0.3090 (-0.0132)**
ResNet50+Long+CutMix (Yun et al., 2019)	76.874/93.500	0.2999	**77.274/93.304**	**0.2890 (-0.0109)**

Empirical Studies

ImageNet-1K (Standard):

Teacher	w/ LS	Acc. (Top1/Top5)	Student	Acc. (Top1/Top5)
ResNet-50	✘	$76.056 \pm 0.119/92.791 \pm 0.106$	ResNet-18	$71.425 \pm 0.038/90.185 \pm 0.075$
			ResNet-50	$76.325 \pm 0.068/92.984 \pm 0.043$
	✔	$\mathbf{76.128 \pm 0.069/92.977 \pm 0.030}$	ResNet-18	$\mathbf{71.816 \pm 0.017/90.466 \pm 0.074}$
			ResNet-50	$\mathbf{77.052 \pm 0.030/93.376 \pm 0.015}$
ResNet-101	✘	77.374/93.546	ResNet-50	77.428/93.712
			ResNet-101	78.270/94.152
	✔	**77.836/93.662**	ResNet-50	**77.624/93.862**
			ResNet-101	**78.476/94.008**

CUB200-2011 (Fine-grained):

Teacher	w/ LS	Acc. (Top1/Top5)	Student	Acc. (Top1/Top5)
ResNet-50	✘	$79.931 \pm 0.037/94.370 \pm 0.064$	ResNet-18	$77.116 \pm 0.086/93.241 \pm 0.108$
			ResNet-50	$80.910 \pm 0.033/94.738 \pm 0.114$
	✔	$\mathbf{81.497 \pm 0.035/95.043 \pm 0.112}$	ResNet-18	$\mathbf{78.382 \pm 0.099/93.621 \pm 0.120}$
			ResNet-50	$\mathbf{82.355 \pm 0.050/95.440 \pm 0.075}$
ResNet-101	✘	80.380/94.491	ResNet-50	81.261/94.905
			ResNet-101	81.572/**95.371**
	✔	**82.332/94.970**	ResNet-50	**82.263/95.320**
			ResNet-101	**82.522**/95.199

iMaterialist-2019_P (Noisy):

Teacher	w/ LS	Acc. (Top1/Top3)	Student	Acc. (Top1/Top3)
ResNet-50	✘	66.241/91.015	ResNet-18	65.250/90.243
			ResNet-50	67.420/92.155
	✔	**66.825/91.669**	ResNet-18	**65.359/90.530**
			ResNet-50	**67.528/92.551**
ResNet-101	✘	66.726/91.263	ResNet-50	67.905/92.481
			ResNet-101	68.281/92.580
	✔	**67.370/91.877**	ResNet-50	**67.925/92.789**
			ResNet-101	**68.618/92.907**

Empirical Studies

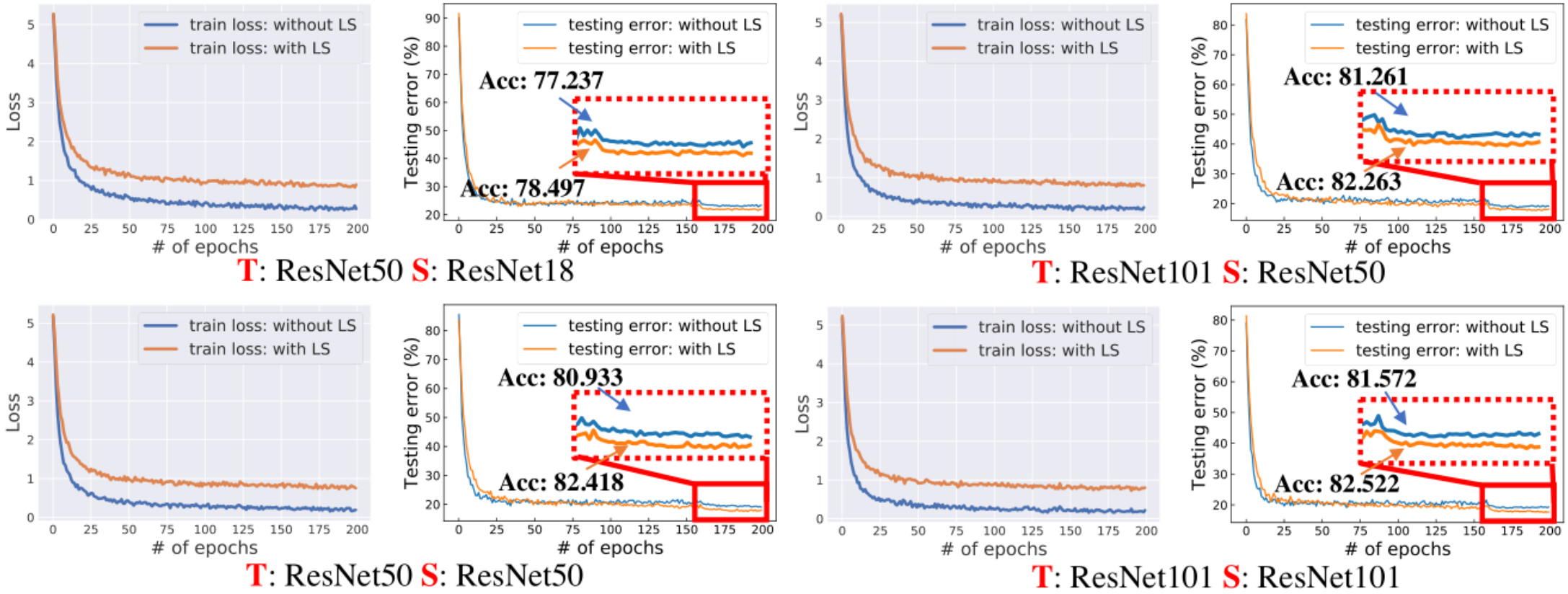
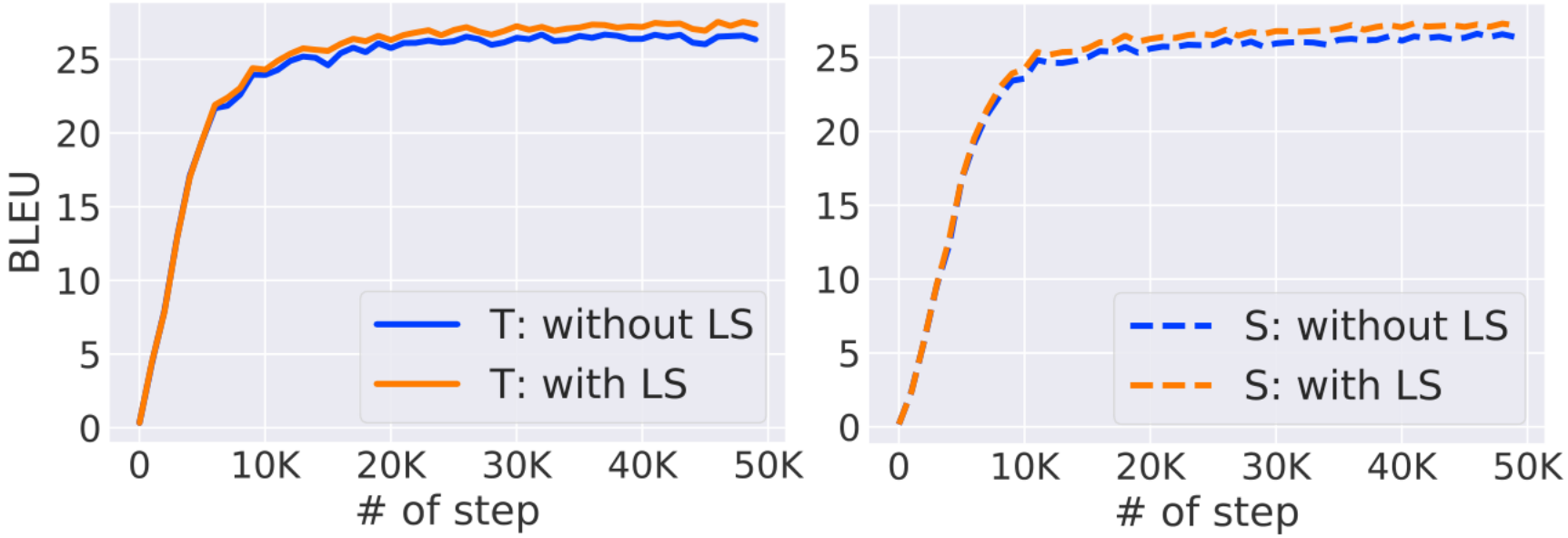


Figure 5: The training and testing curves of knowledge distillation on CUB200-2011 when teachers are trained w/ and w/o label smoothing. The specific teacher and student architectures are given below each subfigure, therein, T indicates the teacher architecture and S indicates the student.

Empirical Studies



What is a Better Teacher in Knowledge Distillation?

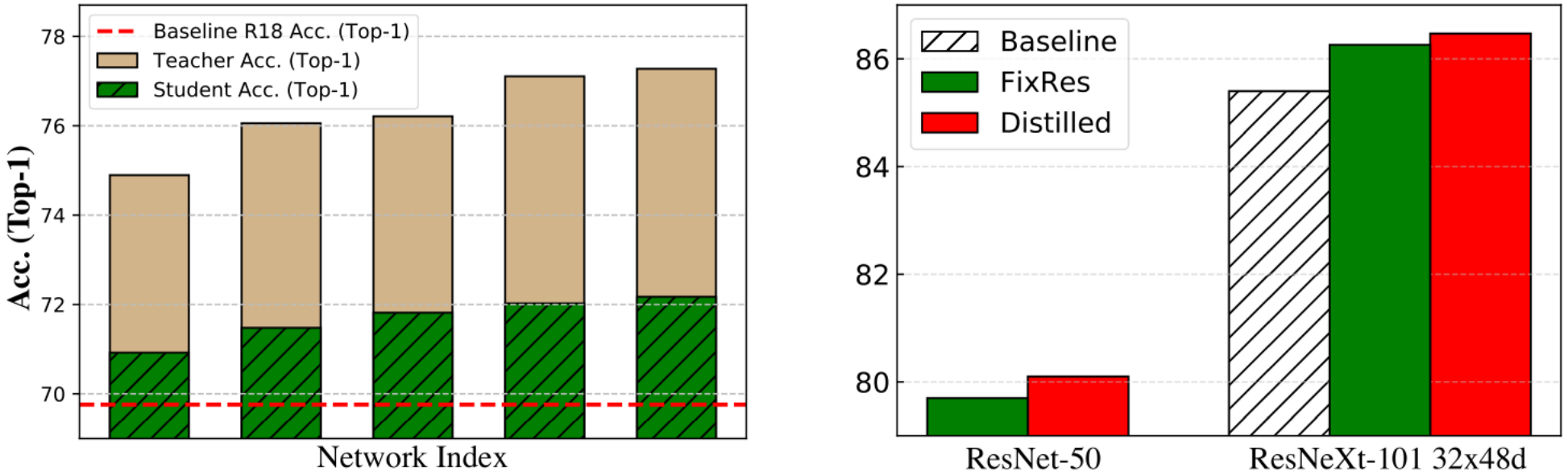


Figure 8: Left is the accuracy relationship between teachers and students, wherein, all teachers are trained with label smoothing. Right is the accuracy of knowledge distillation by using strong teacher to fine-tune the student, FixRes (Touvron et al., 2019) is adopted in both teacher and student networks.

What is a Better Teacher in Knowledge Distillation?

Teacher (same arch)	Acc. (Top-1)	Student	Acc. (Top-1)
R50	76.056	R18	71.478
R50+LS	76.212	R18	71.816
R50+LS+Long	77.106	R18	72.024
R50+LS+Long+CutMix	77.274	R18	72.172

Teacher (different archs)	Acc. (Top-1)	Student	Acc. (Top-1)
MobileNet V2	71.878	R18	70.054
DenseNet-121	74.894	R18	70.922
Wide ResNet-50-2	77.808	R18	72.232
ResNeXt-101-32x8d	79.698	R18	72.412

What is a Better Teacher in Knowledge Distillation?

		ImageNet
Ratio (hard label – soft label)	w/ LS	Acc. (Top1/Top5)
0.3 – ~~0.5~~ 0.7	✘	71.592/90.386
	✔	**71.752/90.412**
0.5 – 0.5	✘	71.484/90.218
	✔	**71.748/90.454**
0.7 – 0.3	✘	71.164/90.196
	✔	**71.314/90.200**

What Circumstances will make LS less Effective?

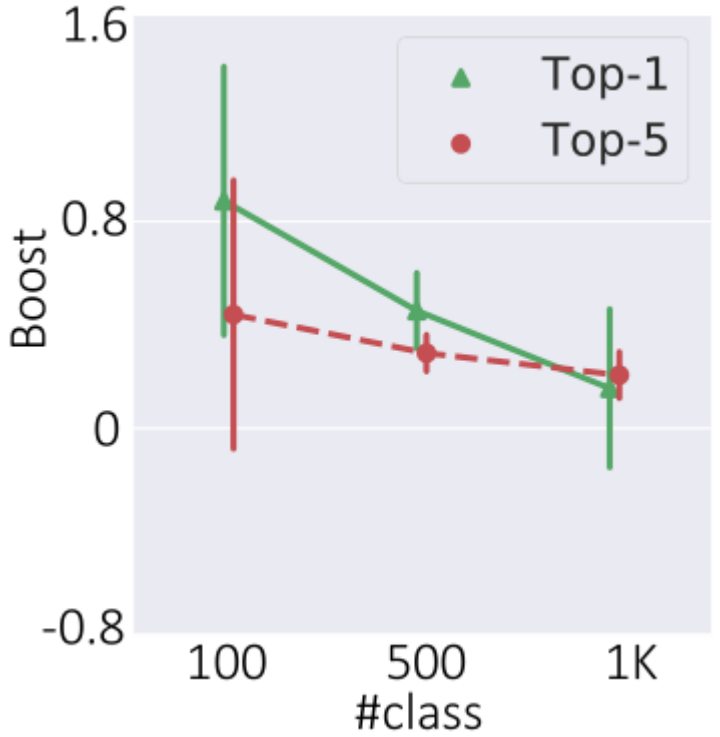
1. Long-Tailed Distribution

Teacher	w/ LS	ImageNet-LT Acc. (Top1/Top5)	Place-LT Acc. (Top1/Top3)	iNaturalist 2019 Acc. (Top1/Top3)
ResNet-18	✘	**39.975/64.645**	**26.479/47.233**	**67.195/83.465**
	✔	39.115/63.655	25.877/46.260	66.700/83.432
ResNet-34	✘	**41.150/66.205**	**27.329/48.753**	**70.165/86.304**
	✔	40.965/65.850	26.863/48.110	69.406/86.073
ResNet-50	✘	**40.985/66.030**	27.384/**48.740**	**73.729/88.845**
	✔	39.965/65.195	**27.562**/47.945	72.904/87.954
ResNet-101	✘	–/–	**28.096/50.164**	**74.389**/88.416
	✔	–/–	27.466/48.781	73.597/**88.779**

What Circumstances will make LS less Effective?

2. More Classes

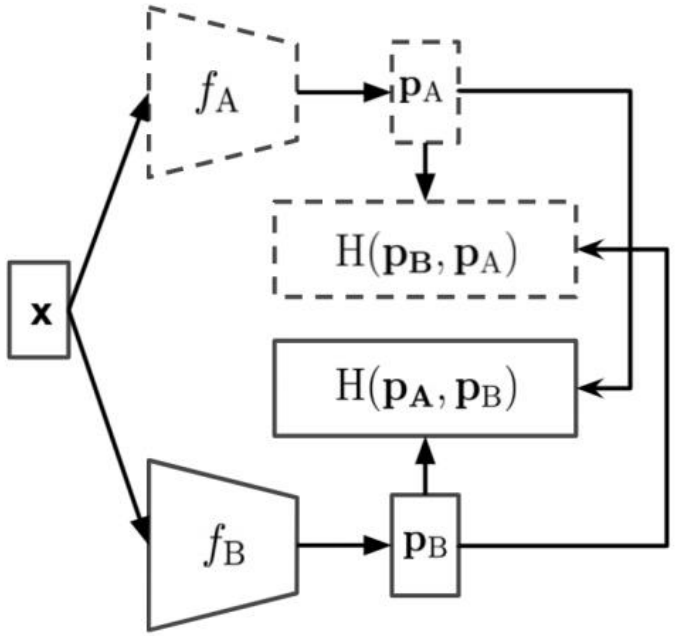
		ImageNet-100	ImageNet-500	ImageNet-1K
Teacher	w/ LS	Acc. (Top1/Top5)	Acc. (Top1/Top5)	Acc. (Top1/Top5)
ResNet-18	✘	82.380/**95.520**	73.521/91.642	**69.758**/89.076
	✔	**82.740**/95.440	**74.123/92.004**	69.606/**89.372**
ResNet-101	✘	82.000/94.340	81.712/95.080	77.374/93.546
	✔	**83.400/95.300**	**82.020/95.300**	**77.836/93.662**
Average (↑)		↑**0.880/0.440**	↑0.455/0.291	↑0.155/0.206



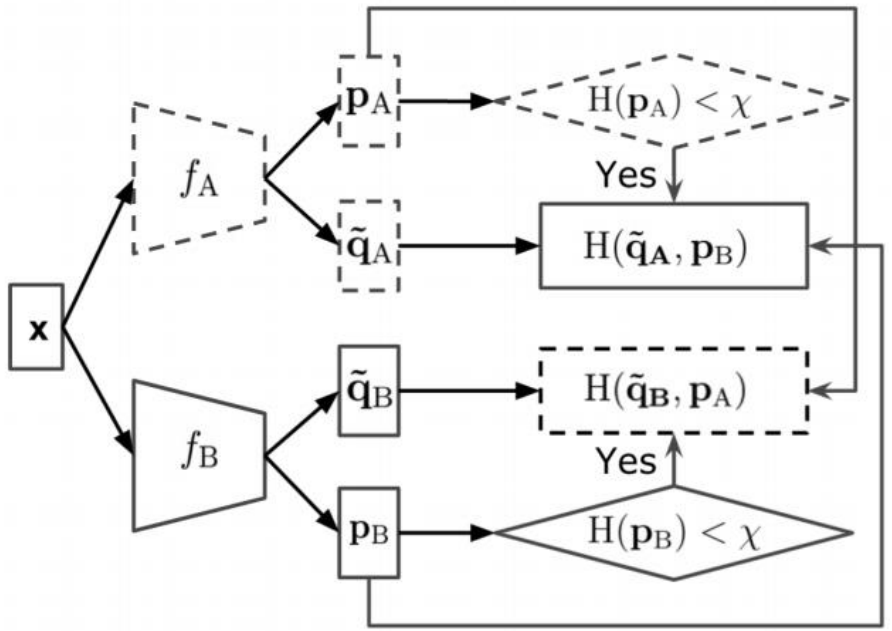
Another paper I want to talk about

Not All Knowledge Is Created Equal

Ziyun Li[1], Xinshao Wang[2,6,*], Haojin Yang[1], Di Hu[3],
Neil M. Robertson[4,6], David A. Clifton[2,5,†], Christoph Meinel[1]



(a) Conventional MKD.



(b) CMD.

$$L_A = L_{A_{SelfKD}} + L_{B2A} = \begin{cases} H(\tilde{q}_A, p_A) + H(\tilde{q}_B, p_A) = 2E_{\frac{\tilde{q}_A + \tilde{q}_B}{2}}(-\log\ p_A), & H(p_B) < \chi, \\ H(\tilde{q}_A, p_A) = E_{\tilde{q}_A}(-\log\ p_A), & H(p_B) \geq \chi. \end{cases}$$

$$L_B = L_{B_{SelfKD}} + L_{A2B} = \begin{cases} H(\tilde{q}_B, p_B) + H(\tilde{q}_A, p_B) = 2E_{\frac{\tilde{q}_A + \tilde{q}_B}{2}}(-\log\ p_B), & H(p_A) < \chi, \\ H(\tilde{q}_B, p_B) = E_{\tilde{q}_B}(-\log\ p_B), & H(p_A) \geq \chi. \end{cases}$$

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

Thanks for Listening