南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Be Your Own Teacher： Improve the Performance of Convolutional Neural Networks via Self Distillation

Linfeng Zhang
Tsinghua University
zhanglinfeng1997@outlook.com

Jiebo Song
IIISCT
songjb@iiisct.com

Anni Gao
IIISCT
gaoan@iiisct.com

Jingwei Chen
Hisilicon
jean.chenjingwei@hisilicom

Chenglong Bao
Tsinghua University
clbao@mail.tsinghua.edu.cn

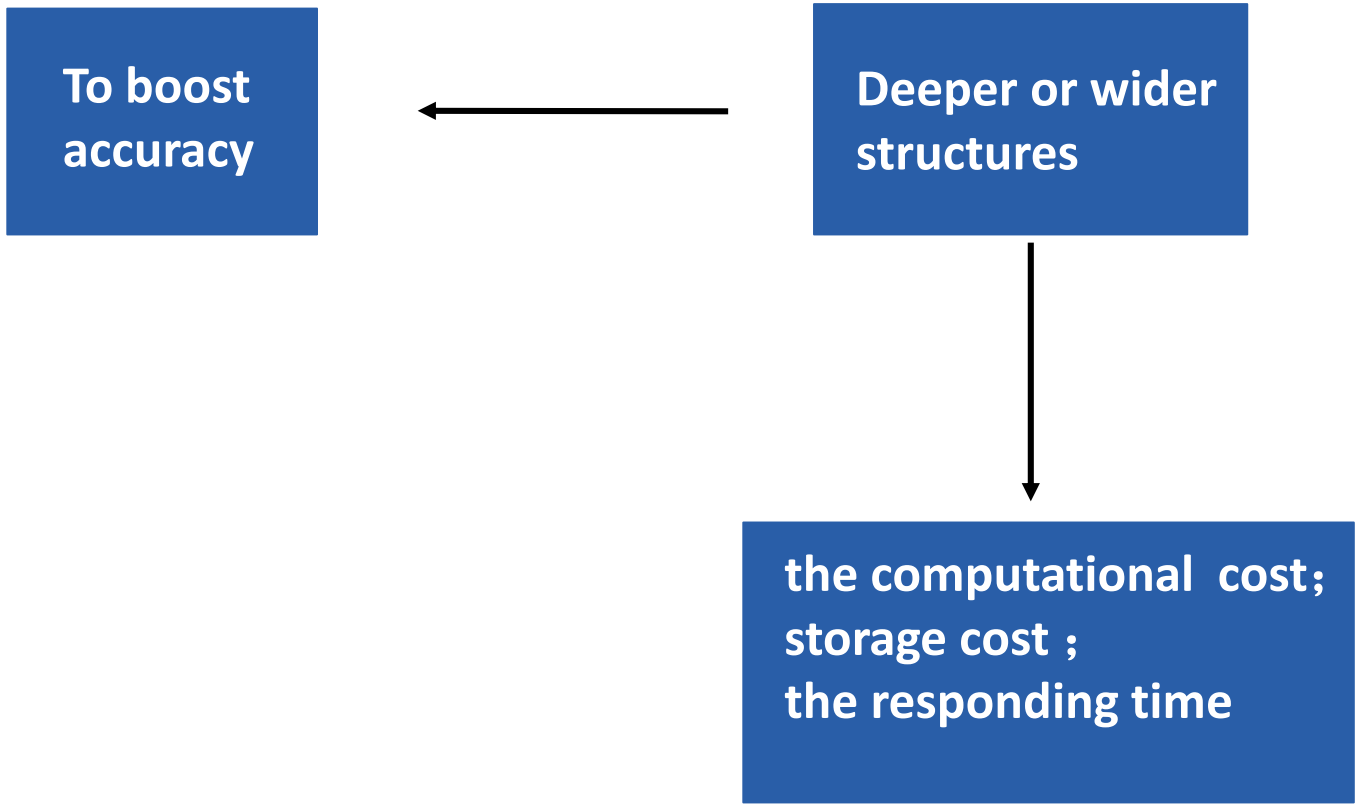
Kaisheng Ma
Tsinghua University
kaisheng@mail.tsinghua.edu.cn

ICCV 2019

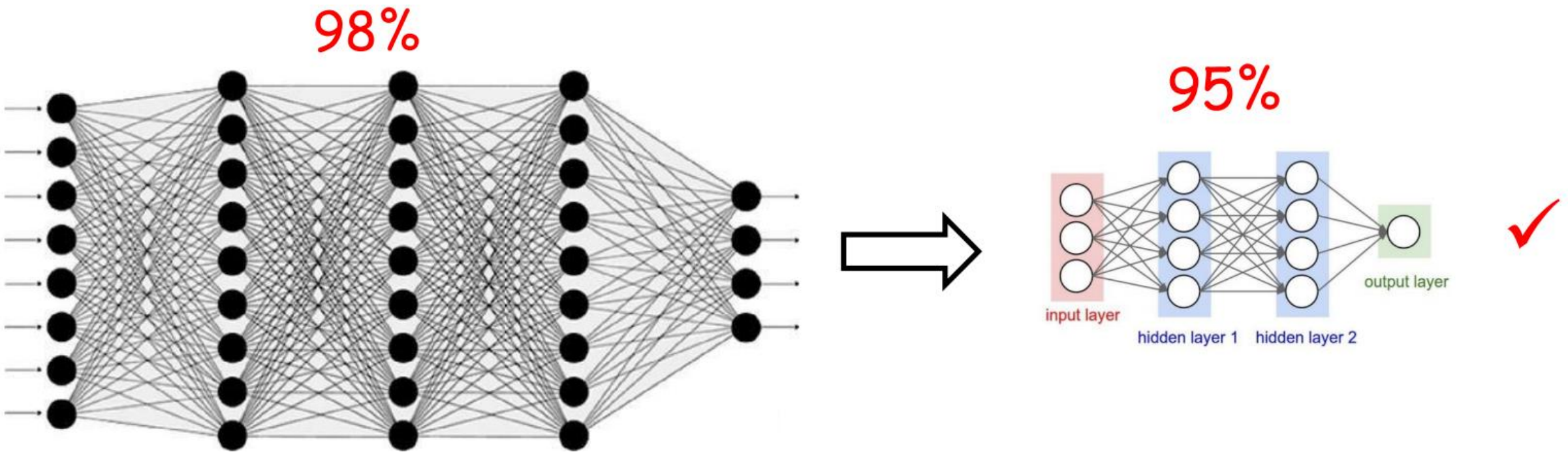
Contents

Motivation

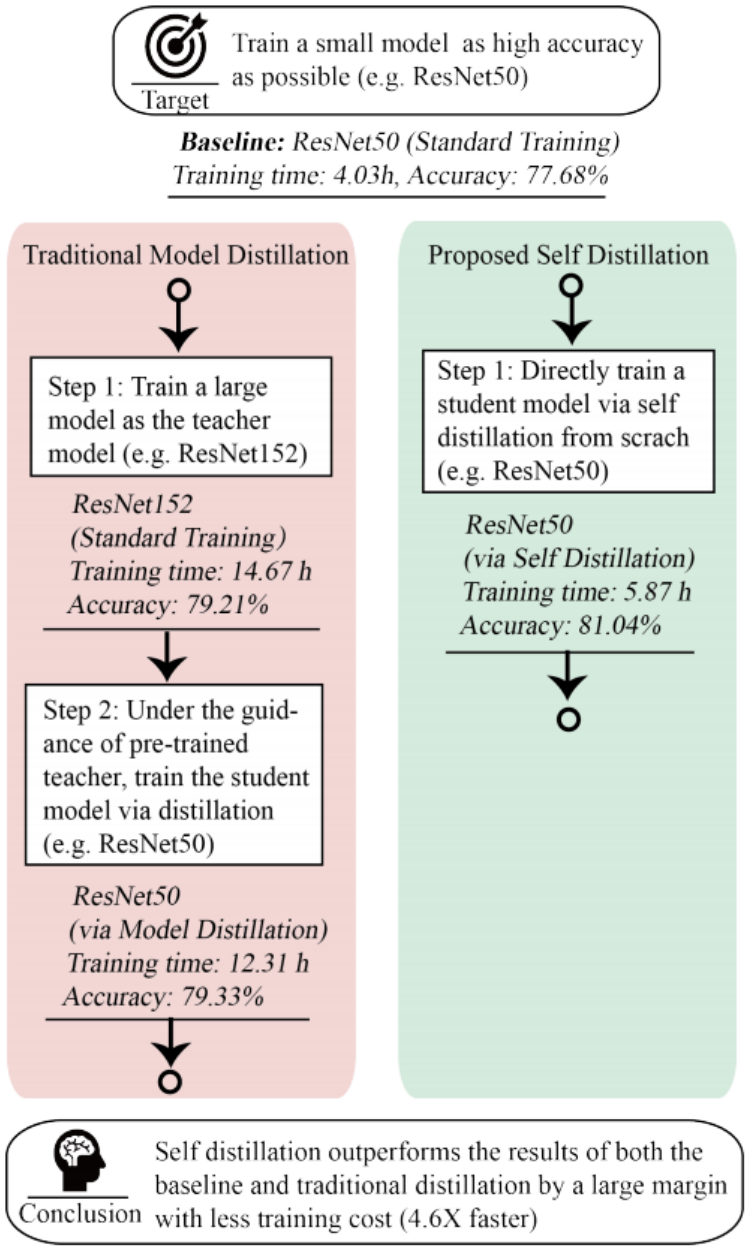
Convolutional neural networks



Knowledge Distillation



Self distillation



Traditional distillation：

to orientate compact student models to approximate over-parameterized teacher models.

Problem：

(1) low efficiency on knowledge transfer;

(2) how to design and train proper teacher models? --- long time

Figure 1. Comparison of training complexity, training time, and accuracy between traditional distillation and proposed self distillation (reported on CIFAR100).

Method

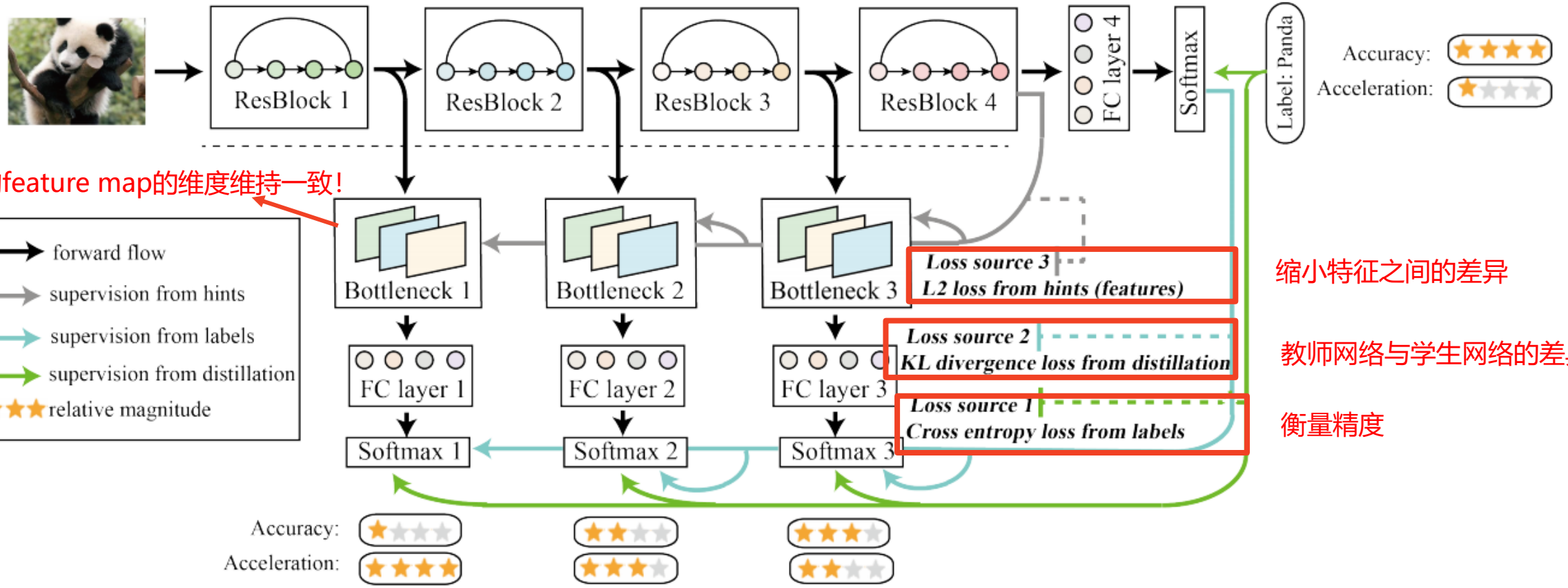
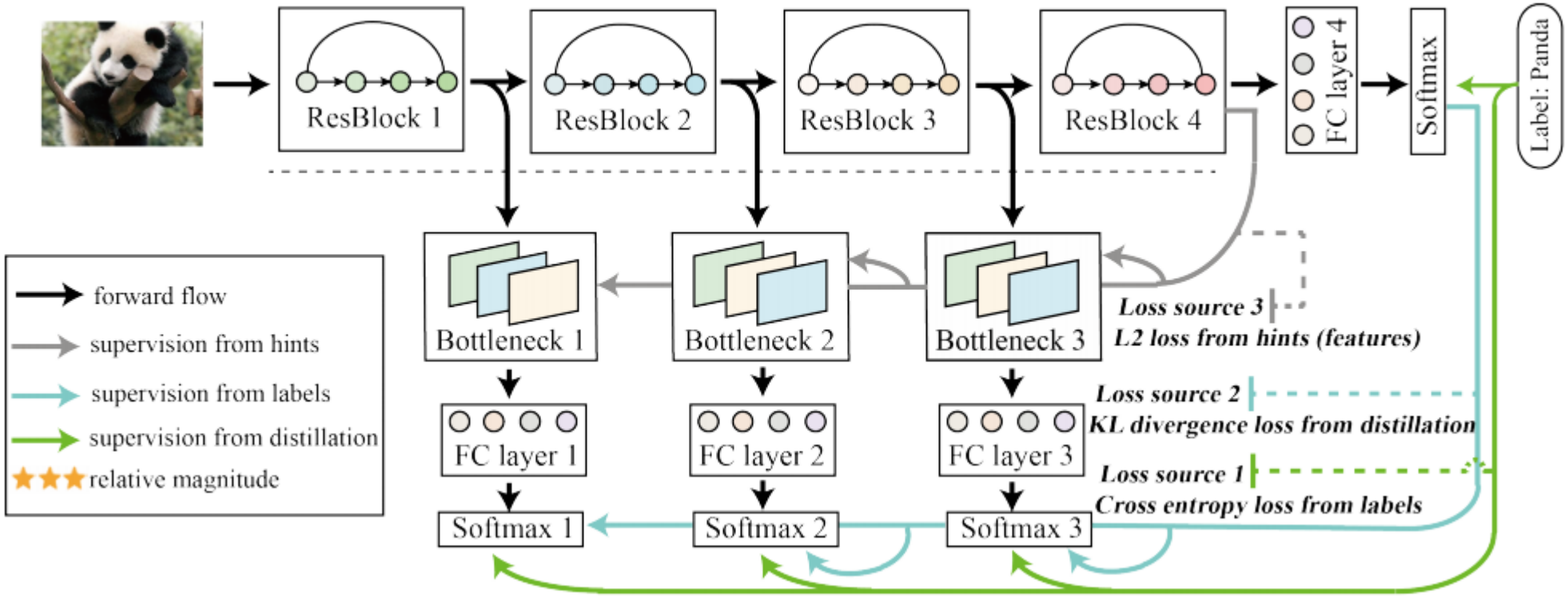


Figure 2. This figure shows the details of a ResNet equipped with proposed self distillation. (i) A ResNet has been divided into four sections according to their depth. (ii) Additional bottleneck and fully connected layers are set after each section, which constitutes multiple classifiers. (iii) All of the classifiers can be utilized independently, with different accuracy and response time. (iv) Each classifier is trained under three kinds of supervision as depicted. (v) Parts under the dash line can be removed in inference.

Self Distillation



$$q_i^c = \frac{\exp\left(z_i^c/T\right)}{\Sigma_j^c \exp\left(z_j^c/T\right)}$$

- Loss Source 1: $(1-\alpha) \cdot CrossEntropy(q^i, y)$
- Loss Source 2: $\alpha \cdot KL(q^i, q^C)$
- Loss Source 3: $\lambda \cdot \|F_i - F_C\|_2^2$

$$loss = \sum_i^C loss_i$$

$$= \sum_i^C \Big((1-\alpha) \cdot CrossEntropy(q^i, y) + \alpha \cdot KL(q^i, q^C) + \lambda \cdot ||F_i - F_C||_2^2 \Big)$$

Experiments---Compared with Standard Training

Five convolutional neural networks:

ResNet, WideResNet , Pyramid , ResNet , ResNeXt , VGG

Two datasets:

CIFAR100 , ImageNet

Neural Networks	Baseline	Classifier 1/4	Classifier 2/4	Classifier3/4	Classifier 4/4	Ensemble
VGG19(BN)	64.47	63.59	67.04	68.03	67.73	68.54
ResNet18	77.09	67.85	74.57	78.23	78.64	79.67
ResNet50	77.68	68.23	74.21	75.23	80.56	81.04
ResNet101	77.98	69.45	77.29	81.17	81.23	82.03
ResNet152	79.21	68.84	78.72	81.43	81.61	82.29
ResNeXt29-8	81.29	71.15	79.00	81.48	81.51	81.90
WideResNet20-8	79.76	68.85	78.15	80.98	80.92	81.38
WideResNet44-8	79.93	72.54	81.15	81.96	82.09	82.61
WideResNet28-12	80.07	71.21	80.86	81.58	81.59	82.09
PyramidNet101-240	81.12	69.23	78.15	80.98	82.30	83.51

Table 1. Experiments results of accuracy (%) on CIFAR100 (the number marked in red is lower than its baseline).

Compared with Distillation

Neural Networks	Baseline	Classifier 1/4	Classifier 2/4	Classifier 3/4	Classifier 4/4	Ensemble
VGG19(BN)	70.35	42.53	55.85	71.07	72.45	73.03
ResNet18	68.12	41.26	51.94	62.29	69.84	68.93
ResNet50	73.56	43.95	58.47	72.84	75.24	74.73

Table 2. Experiments results of top-1 accuracy (%) on ImageNet (the number marked in red is lower than its baseline).

Teacher Model	Student Model	Baseline	KD [15]	FitNet [32]	AT [42]	DML [43]	Our approach
ResNet152	ResNet18	77.09	77.79	78.21	78.54	77.54	78.64
ResNet152	ResNet50	77.68	79.33	80.13	79.35	78.31	80.56
WideResNet44-8	WideResNet20-8	79.76	79.80	80.48	80.65	79.91	80.92
WideResNet44-8	WideResNet28-12	80.07	80.95	80.53	81.46	80.43	81.58

Table 3. Accuracy (%) comparison with traditional distillation on CIFAR100.

Compared with Deeply Supervised Net

Neural Networks	Method	Classifier 1/4	Classifier 2/4	Classifier3/4	Classifier 4/4	Ensemble
ResNet18	DSN	67.23	73.80	77.75	78.38	79.27
	Our approach	67.85	74.57	78.23	78.64	79.67
ResNet50	DSN	67.87	73.80	74.54	80.27	80.67
	Our approach	68.23	74.21	75.23	80.56	81.04
ResNet101	DSN	68.17	75.43	80.98	81.01	81.72
	Our approach	69.45	77.29	81.17	81.23	82.03
ResNet152	DSN	67.60	77.04	81.06	81.35	81.83
	Our approach	68.84	78.72	81.43	81.61	82.29

Table 4. Accuracy (%) comparison with deeply supervised net [24] on CIFAR100.

(i) Self distillation outperforms deep supervision in every classifier.
(ii) Shallow classifiers benefit more from self distillation

Discussion--- flat minima

The possible explanations of notable performance improvement?

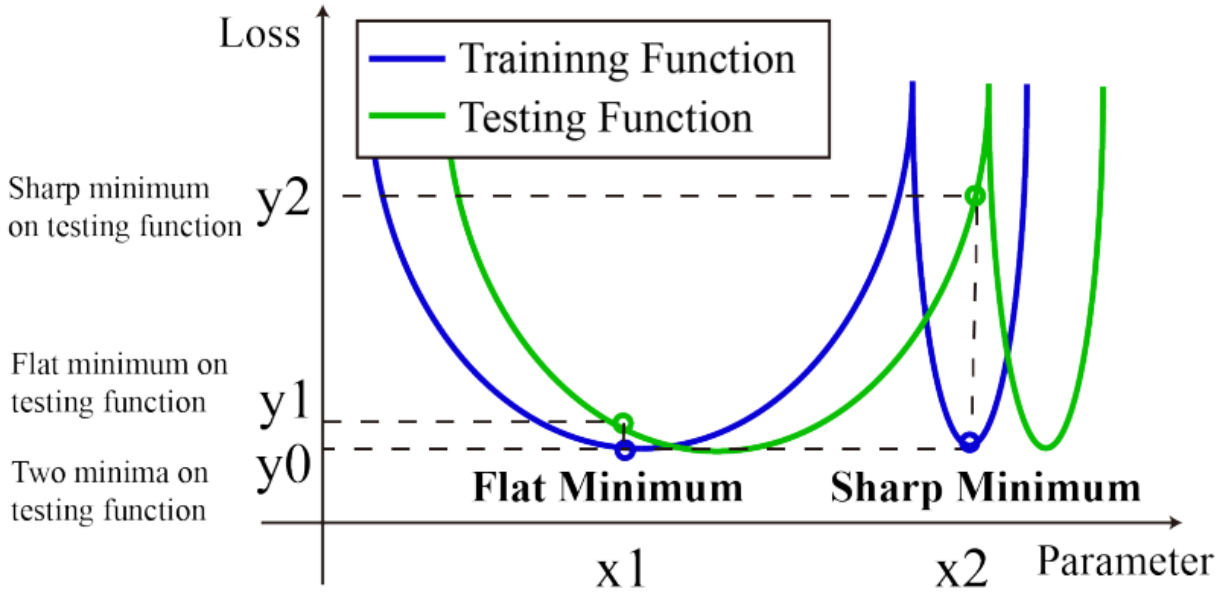
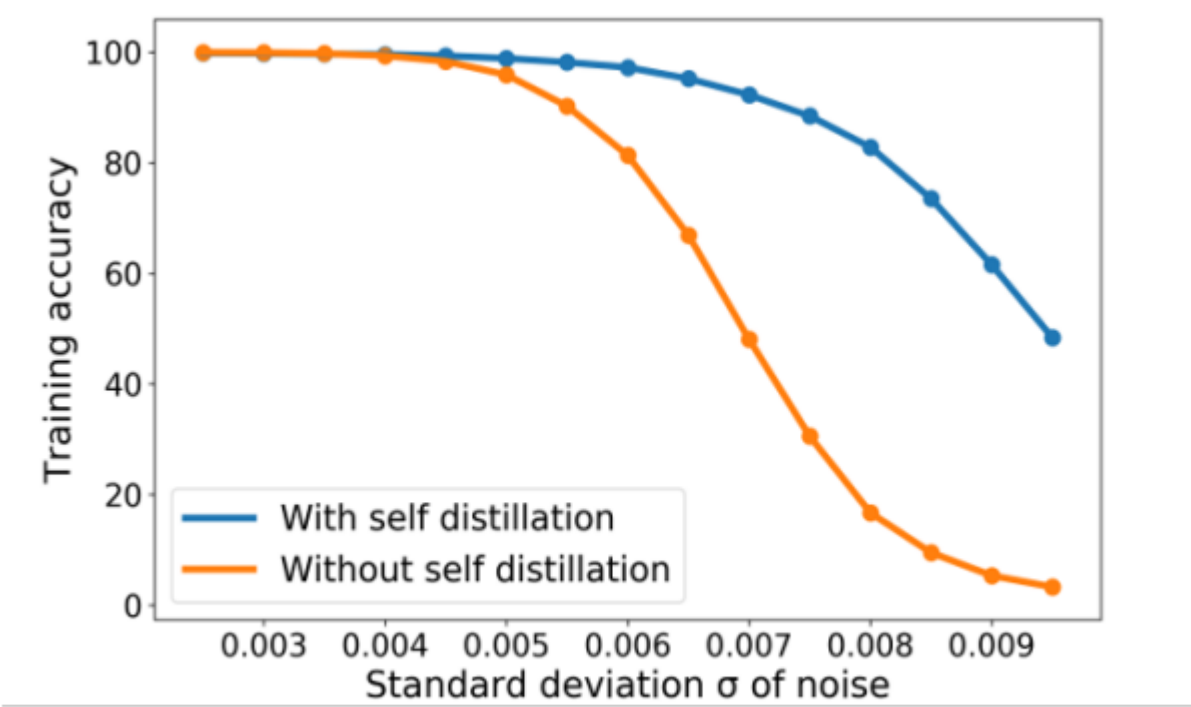


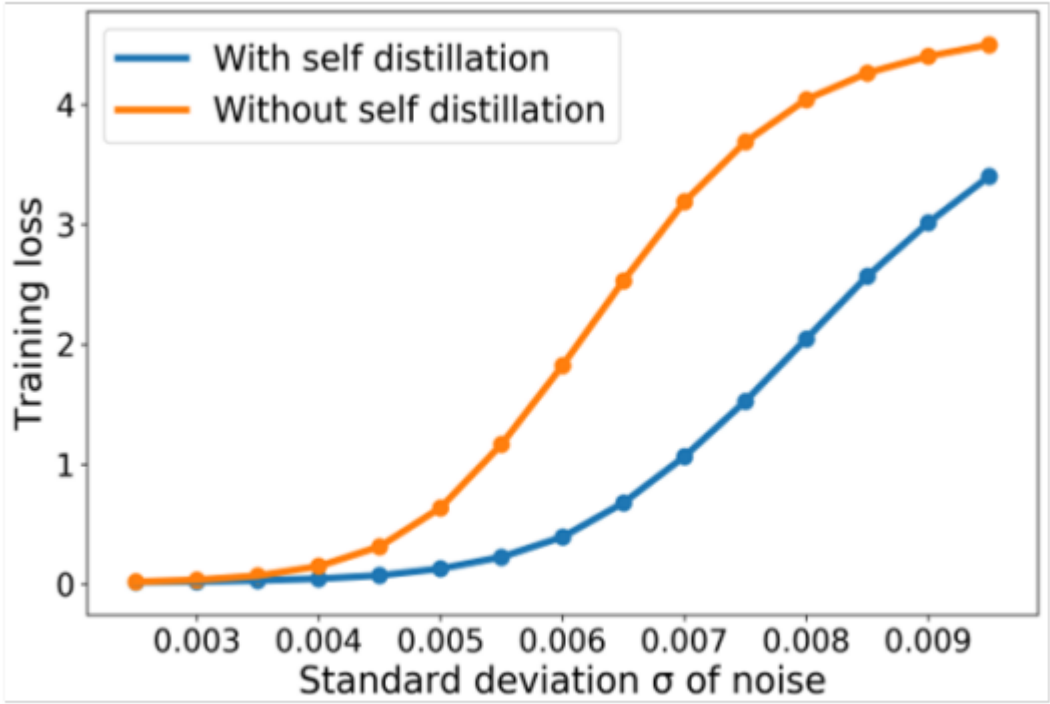
Figure 3. An intuitive explanation of the difference between flat and sharp minima [20].

Discussion--- flat minima

The possible explanations of notable performance improvement?



(a) Training accuracy



(b) Training loss

Figure 4. Comparison of training accuracy and loss with increasing Gaussian noise:models trained with self distillation are more tolerant to noise - flat minima.

Discussion--- vanishing gradients

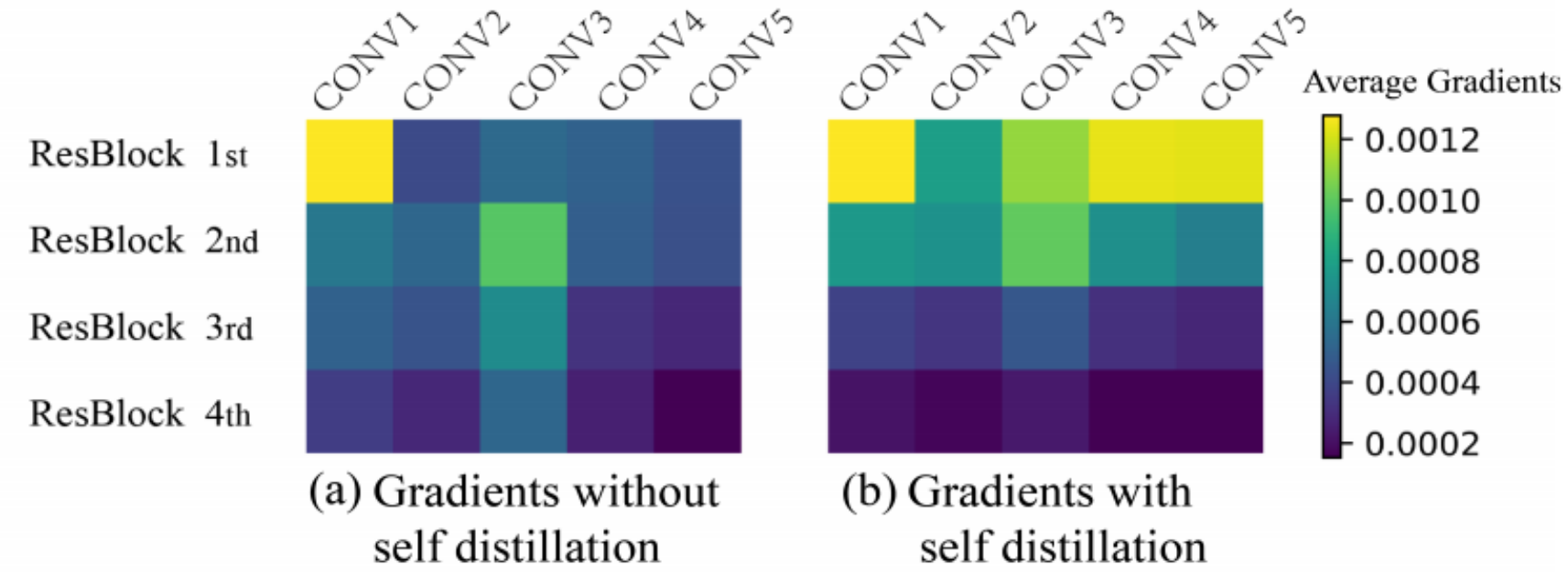
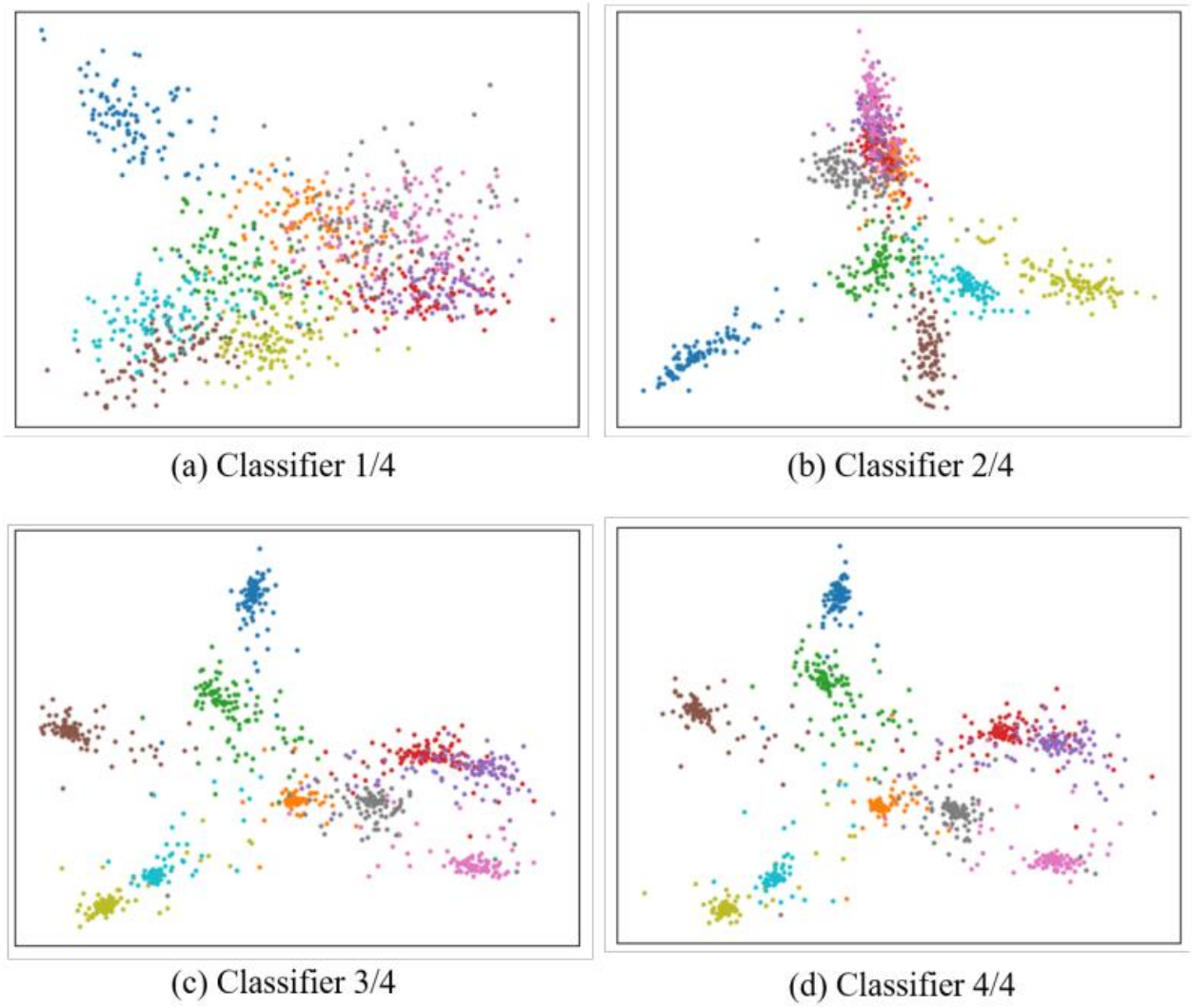


Figure 5. Statistics of layer-wised gradients.

Discussion--- discriminating features



(a) Classifier 1/4

(b) Classifier 2/4

(c) Classifier 3/4

(d) Classifier 4/4

Figure 6. PCA (principal component analysis) visualization of feature distribution in four classifiers.

Conclusion

- 提出的自蒸馏训练框架能够大大的增加训练后模型的精度。
- 相比于传统的蒸馏方法，提出了一段式的蒸馏方法，将模型的训练时间缩短。
- 相比于其他改进的蒸馏方法，模型的精度得到提升。
- 不仅可以提升模型的精度，还可以在一定精度的要求下，对模型的结构进行裁剪。

THANKS