Contrastive Deep Supervision

Linfeng Zhang, Xin Chen, Junbo Zhang, Runpei Dong, Kaisheng Ma

ECCV TEL AVIV 2022

Background

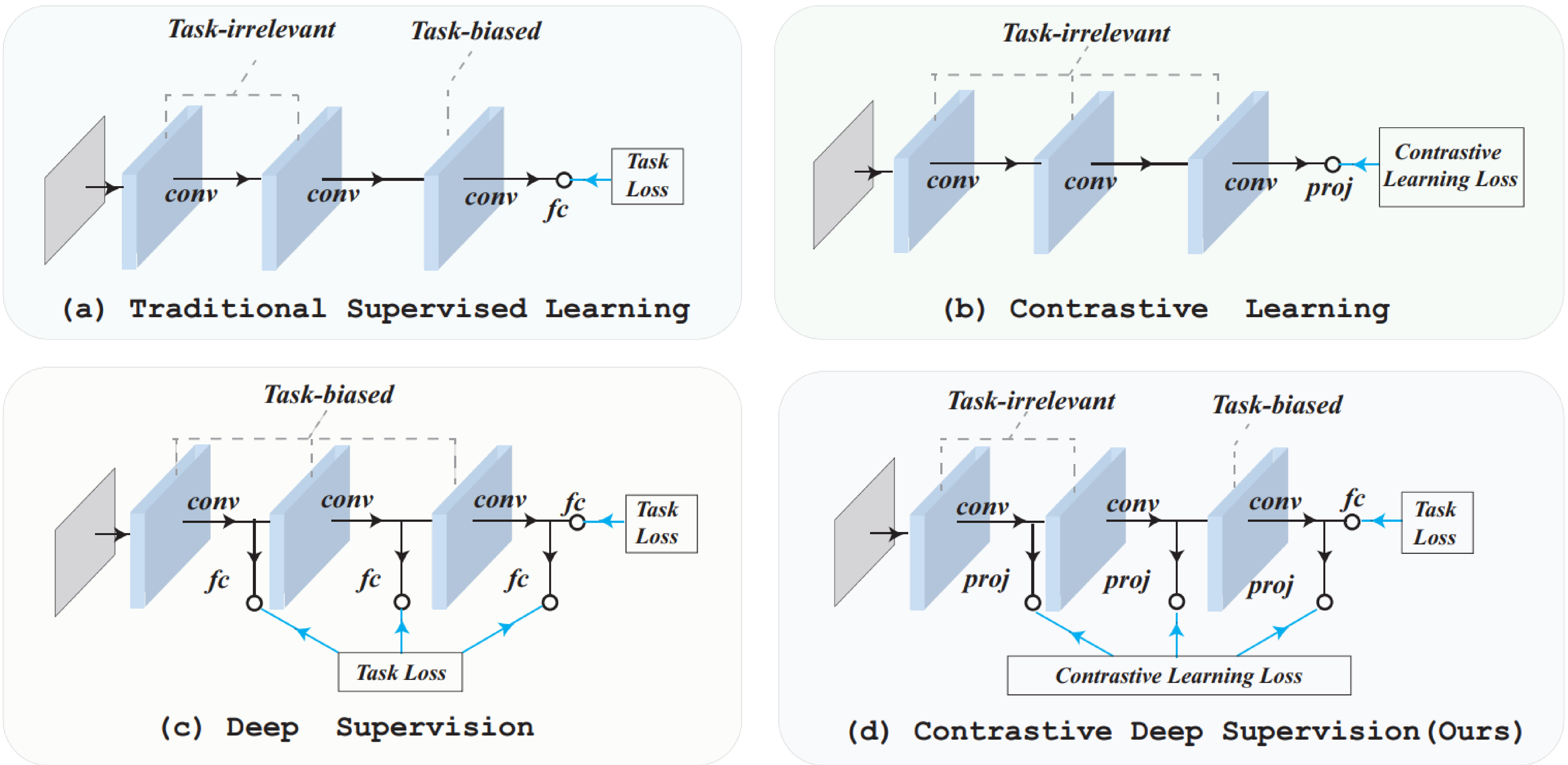


Fig. 1. The overview of the four methods. "→" and "→" indicate the path of forward computation and gradients backward computation. "proj" and "fc" indicate the projection heads and the fully connected classifiers, respectively. The gray dash line indicates whether the feature is task-irrelevant or task-biased. (a) Traditional supervised learning only applies supervision to the last layer and propagates it to the previous layers, leading to gradient vanishing. (c) Deep supervision trains both the last layer and the intermediate layers directly, which addresses gradient vanishing but makes all the layers be biased to the task. (d) Our method introduces contrastive learning to supervise the intermediate layer and thus avoid these problems.

Methodology

Instead of supervising the intermediate layers with the task loss, we propose to supervise them **with Contrastive Learning loss** (InfoNCE), where positive & negative pairs are built with both labels and data augmentation. It can be formulated as

$$\mathcal{L}_{\text{Contra}} = -\sum_{i=1}^{N} \log \frac{\exp(z_i \cdot z_{i+N})/\tau}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(z_i \cdot z_k)/\tau}, \quad \mathcal{L}_{\text{CDS}} = \underbrace{\mathcal{L}_{\text{CE}}(c_K(\mathcal{X}), \mathcal{Y})}_{\text{from standard training}} + \lambda_1 \underbrace{\sum_{i=1}^{K-1} \mathcal{L}_{\text{Contra}}(\mathcal{X}; c_i)}_{\text{from our method}},$$

Experiment

We have evaluated our method on general image classification, fine-grained image classification, object detection, in supervised learning, semi-supervised learning and knowledge distillation learning,

Table 3. Comparison with the other deep supervision methods on ImageNet.

Metric	Model	Baseline	DSN	DKS	DHM	Ours
top-1	RNT18	69.21	69.54	71.32	71.29	**72.85**
	RNT34	73.17	73.29	74.01	73.89	**76.19**
	RNT50	75.30	75.37	76.47	76.57	**78.25**
top-5	RNT18	89.01	88.87	89.20	90.06	**91.30**
	RNT34	91.24	91.30	91.87	91.66	**93.08**
	RNT50	92.20	92.49	93.60	93.24	**93.99**

3.20% improvements on average

Table 7. Comparison experiments (top-1 and top-5 accuracy / %) with the other eight knowledge distillation methods on ImageNet with ResNet. Numbers in bold indicate the highest. Results marked with † come from the paper of SSKD [67].

Metric	Model	Base	KD	AT	RKD	SP	CRD	CC†	OKD†	SSKD†	Ours
top-1	RNT18	69.21	70.52	70.74	70.63	70.61	71.07	69.96	70.55	71.62	**73.23**
	RNT34	73.17	74.44	74.69	74.61	74.60	74.99	–	–	–	**76.65**
	RNT50	75.30	76.62	76.79	76.92	76.88	77.21	–	–	–	**78.68**
top-5	RNT18	89.01	89.88	90.00	89.71	89.80	91.06	89.17	89.59	90.67	**91.56**
	RNT34	91.24	92.07	92.18	92.14	92.10	92.58	–	–	–	**93.38**
	RNT50	92.20	93.36	93.51	93.60	93.58	93.88	–	–	–	**94.42**

3.62% improvements on average

Method	Batchsize	Epoch	AutoAug	top-1 acc. (%)
Baseline[1]	256	90	×	75.3
Baseline[2]	4096	270	✓	77.6
SupCon[1]	6144	350	✓	78.7
SupCon[2]	512	350	✓	74.5
SupCon[3]	6144	100	✓	77.0
BYOL	1024	1080	×	77.7
BYOL+DSN	1024	1080	×	78.2
Ours[1]	256	90	×	78.3
Ours[2]	256	90	×	78.7
Ours[3]	256	350	✓	79.8

Outperforming BYOL+DSN by **1.6%**

Table 6. Comparison (top-1 acc. %) with deep supervision methods with ResNet50 for fine-grained classification. Models are finetuned from ImageNet pre-trained weights.

Method	CUB	Cars	Flowers	Dogs	Aircrafts
Baseline	78.50	90.25	97.68	76.47	87.43
DSN	$80.14_{+1.64}$	$91.32_{+1.07}$	$98.64_{+0.96}$	$77.21_{+0.74}$	$89.31_{+1.88}$
DKS	$81.34_{+2.84}$	$92.54_{+2.29}$	$99.01_{+1.33}$	$78.32_{+1.85}$	$89.20_{+1.77}$
DHM	$81.27_{+2.77}$	$92.31_{+2.06}$	$98.84_{+1.16}$	$78.20_{+1.73}$	$89.57_{+2.14}$
Ours	$\mathbf{82.10}_{+3.60}$	$\mathbf{92.90}_{+2.65}$	$\mathbf{99.39}_{+1.71}$	$\mathbf{80.99}_{+4.52}$	$\mathbf{90.52}_{+3.09}$

3.11% improvements on average

Table 1. Comparison experiments (top-1 accuracy / %) with the other deep supervision methods on CIFAR100.

Method	RNT18	RNT50	RNT101	RXT50	RXT101	WRN50	WRN101	SET18	SET50	PAT18
Base	77.45	77.81	78.65	79.85	80.67	79.46	79.98	77.46	78.02	76.84
DSN	78.30	78.96	79.37	81.02	81.70	80.98	81.30	78.28	79.46	77.40
DKS	78.96	80.95	81.39	82.27	82.98	81.95	82.58	79.32	80.76	78.96
DHM	78.82	81.12	81.27	82.14	83.27	81.76	82.76	79.14	80.72	78.32
Ours	**80.84**	**81.31**	**83.12**	**82.81**	**83.87**	**82.28**	**83.93**	**80.13**	**81.51**	**80.76**

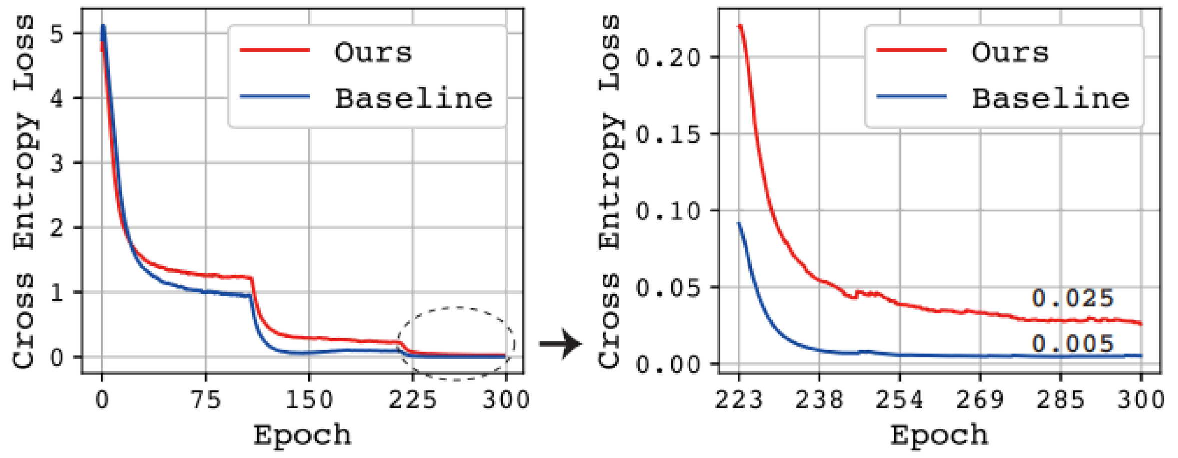
3.44% improvements on average

Table 4. Experiments on different object detection models on COCO2017. ResNet50 models are pre-trained on ImageNet with different deep supervision methods and then utilized as the backbones of these detectors.

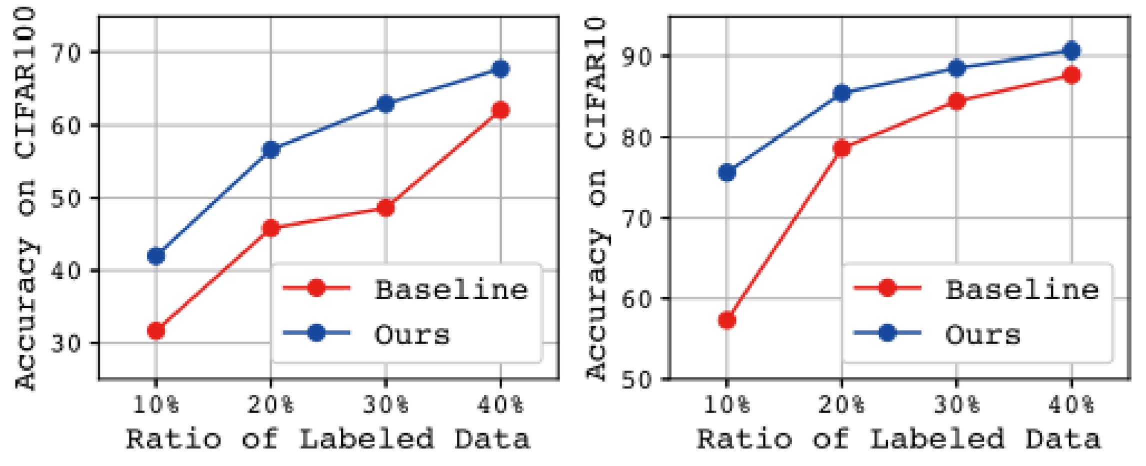
Model	Method	AP	AP_S	AP_M	AP_L
Faster RCNN	Baseline	37.4	21.2	41.0	48.1
	DSN	$37.3_{-0.1}$	$21.0_{-0.2}$	$40.8_{-0.2}$	$48.3_{-0.2}$
	DKS	$37.5_{+0.1}$	$21.2_{+0.0}$	$41.5_{+0.5}$	$47.6_{-0.5}$
	DHM	$37.6_{+0.2}$	$21.3_{+0.1}$	$41.3_{+0.3}$	$48.2_{+0.1}$
	Ours	$\mathbf{38.3}_{+0.9}$	$\mathbf{21.6}_{+0.4}$	$\mathbf{42.0}_{+1.0}$	$\mathbf{50.1}_{+2.0}$
RetinaNet	Baseline	36.5	20.4	40.3	48.1
	DSN	$36.3_{-0.2}$	$20.1_{-0.3}$	$40.0_{-0.3}$	$48.1_{0.0}$
	DKS	$36.7_{+0.2}$	$20.1_{-0.3}$	$40.9_{+0.6}$	$48.2_{+0.1}$
	DHM	$36.7_{+0.2}$	$20.0_{-0.4}$	$40.7_{+0.4}$	$48.5_{+0.4}$
	Ours	$\mathbf{37.3}_{+0.8}$	$\mathbf{21.2}_{+0.8}$	$\mathbf{41.0}_{+0.7}$	$\mathbf{47.9}_{-0.2}$

0.85AP Improve. on average

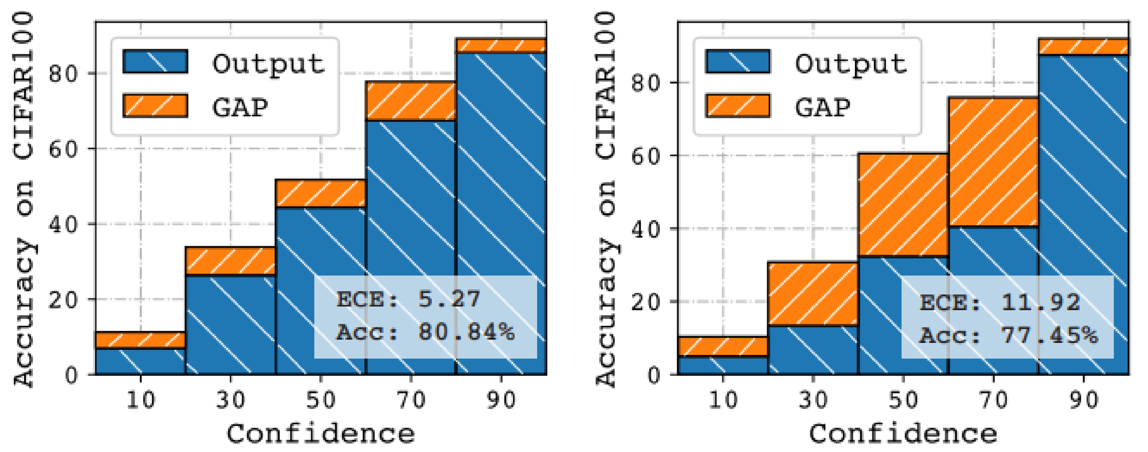
Discussion



With the propose Contrastive Deep Supervision, during the last several training epochs , the model has higher cross entropy loss but higher accuracy, indicating it has effects of regularization.



Contrastive Deep Supervision can also be used in Semi-supervised learning by building positive/negative pairs with only data augmentation



Models trained with our method has better uncertainty estimation on classification tasks.