L2T-DLN: Learning to Teach with Dynamic Loss Network

Zhaoyang Hai

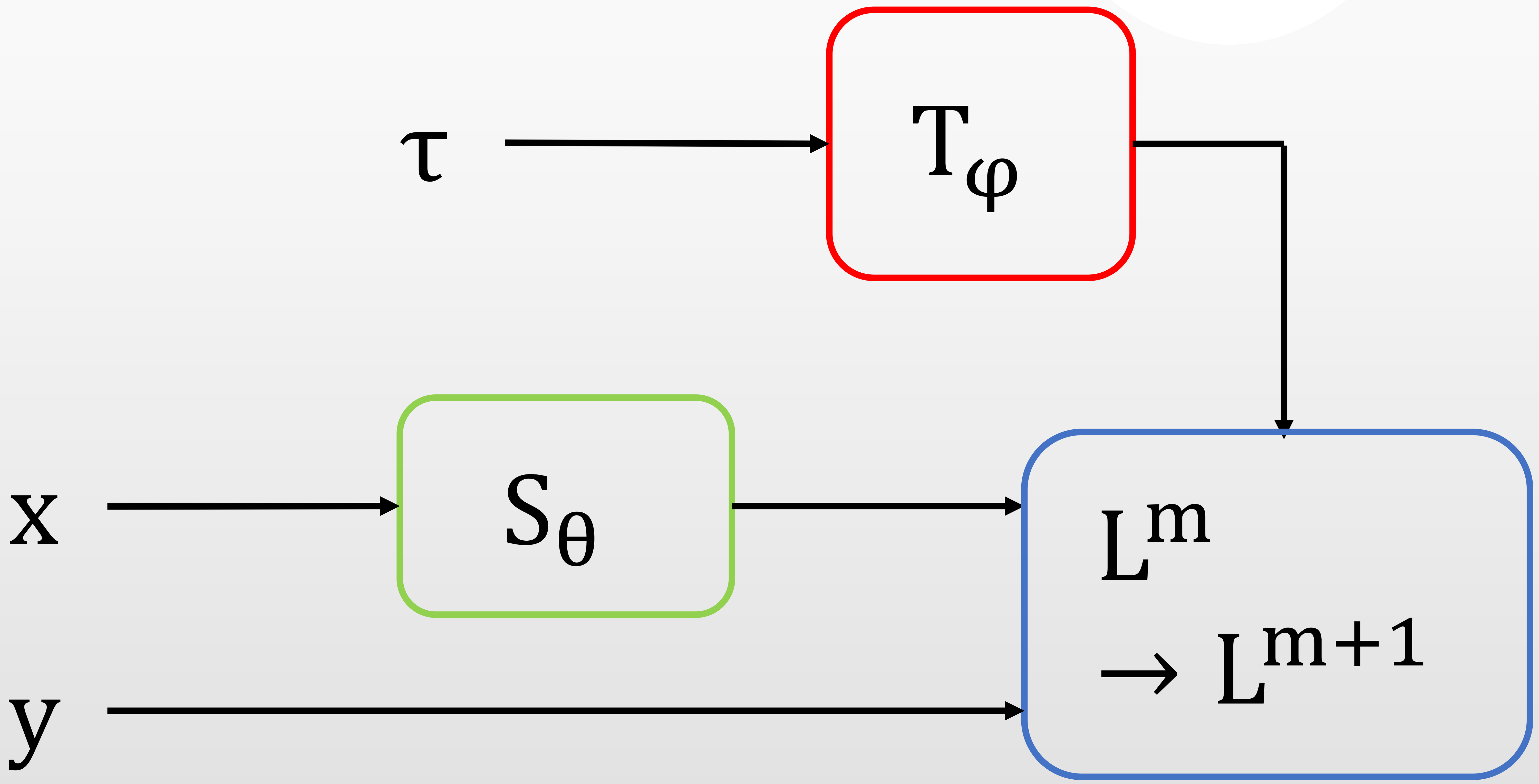
Liyuan Pan

Xiabi Liu

E-mail: {haizhaoyang, Liyuan.pan, liuxiabi}@bit.edu.cn

Teaching-Learning Transaction

Motivation



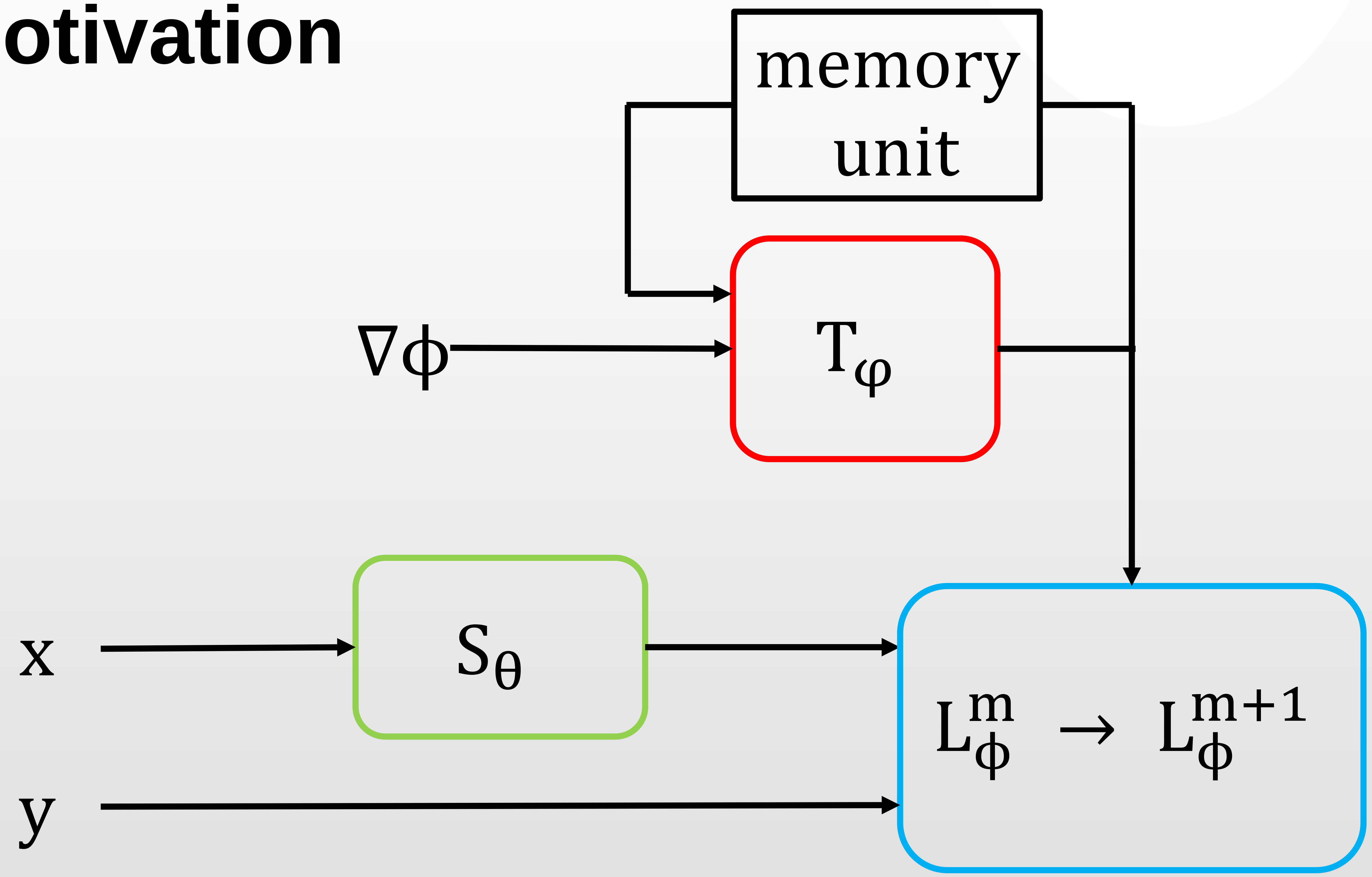
x: training data
y: label
τ: states of student
S_θ: student model
L^m: dynamic loss function
T_φ: teacher model

Two problems in existing works:

1) neglecting the temporal nature of loss function adjustment;

2) neglecting the states of loss functions.

Motivation



x: training data
y: label
$\nabla\phi$: gradient concerning dynamic loss
S_θ: student model
L_ϕ^m: dynamic loss network
T_φ: teacher model

1) adopting an LSTM teacher to accumulate the experience during teaching a student;

2) employing the state of DLN to update the parameter of DLN.

Advantages

Our L2T-DLN bring two benefits:

1) capturing and maintaining short- and long-term dependencies during teaching process;

2) the gradient concerning DLN achieves holistic information integration throughout the learning process, facilitated by prior knowledge (chain rule).

Method

Student learning
θ^0
S_θ^0
…
BP
θ^N
S_θ^N
…
BP
θ^{2N}
S_θ^{2N}
BP
…
$\theta^{(M+K)N}$
DLN learning
ϕ^0
L_ϕ^0
+
ϕ^1
L_ϕ^1
fix
L_ϕ^1
+
…
ϕ^M
$\nabla\phi^0$
g^0
$\nabla\phi^1$
Teacher learning
h^0
T_φ^0
h^1
T_φ^1
φ^0
BP
φ^1
…
φ^K

Convergence Analysis

Conclusion 1: Let $\mathcal{H} \triangleq \nabla^2 e(x)$ denote the Hessian matrix at ϵ-second-order stationary solution v^* where $\lambda_{min}(\mathcal{H}) \leq -\gamma$ and $\gamma > 0$. We have

$$\lambda_{max}(M^{-1}G) > 1 + \eta\gamma/(1 + C/C_{max})$$

Results

Comparison with SOTA loss functions in classification task.

Method	CIFAR-10 ResNet8	CIFAR-10 ResNet20	CIFAR-10 ResNet32	CIFAR-10 WRN	CIFAR-100 ResNet8	CIFAR-100 ResNet20	CIFAR-100 ResNet32	ImageNet NASNet-A	length
CE	87.6	91.3	92.5	96.2	60.2	67.7	69.6	73.5	-
Smooth [7]	87.9	91.5	92.6	96.2	60.5	68.0	69.9	-	-
L-M Softmax [6]	88.7	92.0	93.0	96.3	61.1	68.4	70.4	-	-
L2T-DLF [10]	89.2	92.4	93.1	96.6	61.7	69.0	70.8	-	1
ARLF [1]	89.5	91.5	92.2	95.9	60.2	67.8	69.9	-	-
SLF [5]	89.8	93.0	93.6	**97.1**	62.7	69.9	71.5	-	-
ALA [3]	-	-	93.2	96.7	62.2	69.5	70.9	**74.6**	200
Ours	**90.7** $\pm$ **0.06**	**93.4** $\pm$ **0.18**	**93.8** $\pm$ **0.20**	96.7 $\pm$ 0.09	**63.5** $\pm$ **0.07**	**70.4** $\pm$ **0.03**	**72.0** $\pm$ **0.11**	74.2	25

Comparison with SOTA method in noisy-label classification task.

Method	CIFAR-10 p=20%	CIFAR-10 p=40%	CIFAR-100 p=20%	CIFAR-100 p=40%
Baseline	76.83	70.77	50.86	43.01
MentorNet [4]	86.36	81.76	61.97	52.66
Meta-Weight-Net [9]	90.33	87.54	64.22	58.64
L2R [2]	91.05	88.71	66.08	60.51
Ours	**92.11** $\pm$ **0.27**	**89.39** $\pm$ **1.20**	**70.05** $\pm$ **0.23**	**61.27** $\pm$ **0.51**

Results

Comparison with YOLO-v3 loss in objective detection.

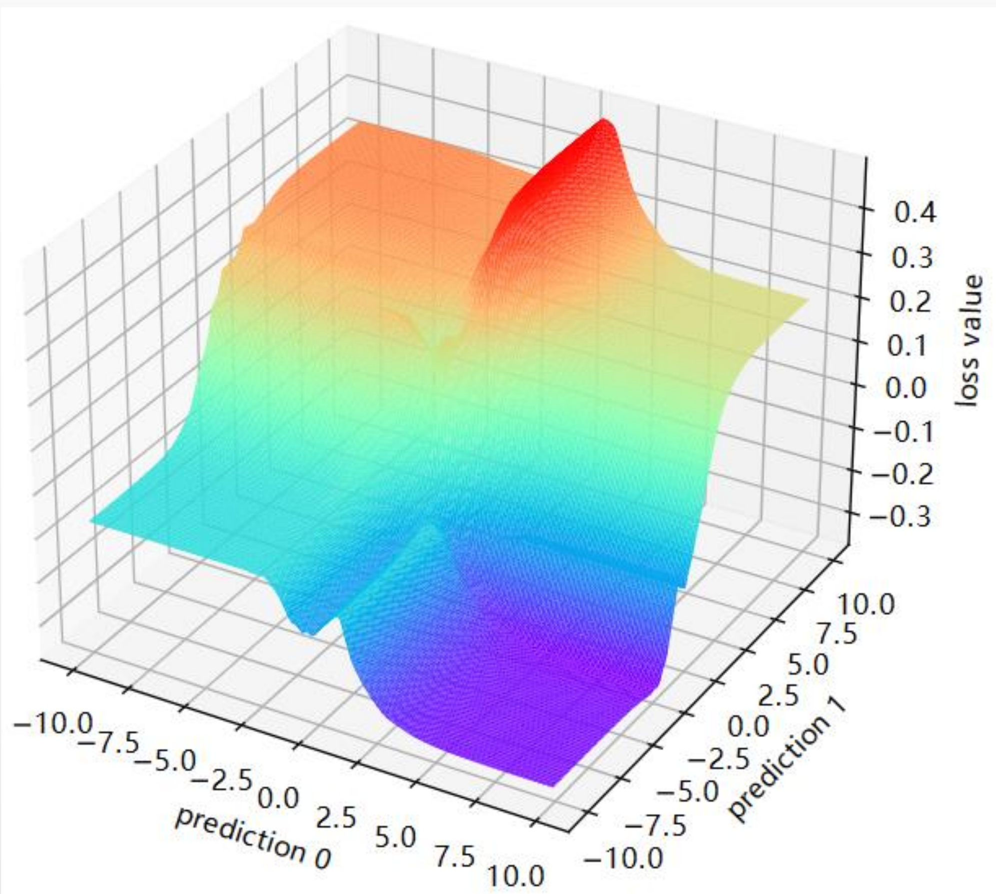
Detectors	Size	mAP	FPS
YOLOV3 [8]	416	55.3	35
YOLOV3-ours	416	56.9	35

Comparison with PSPNet loss in semantic segmentation.

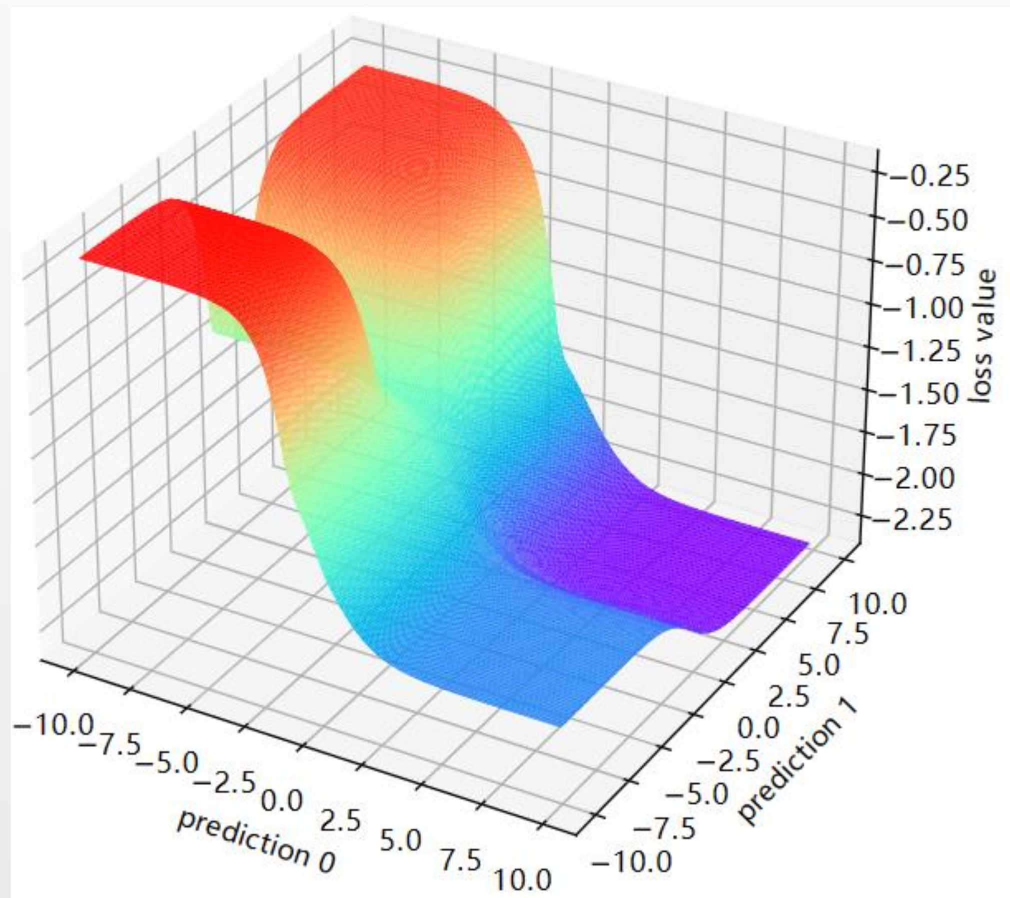
Method	mIoU
PSPNet [11]	82.6
PSPNet-ours	82.9

More details

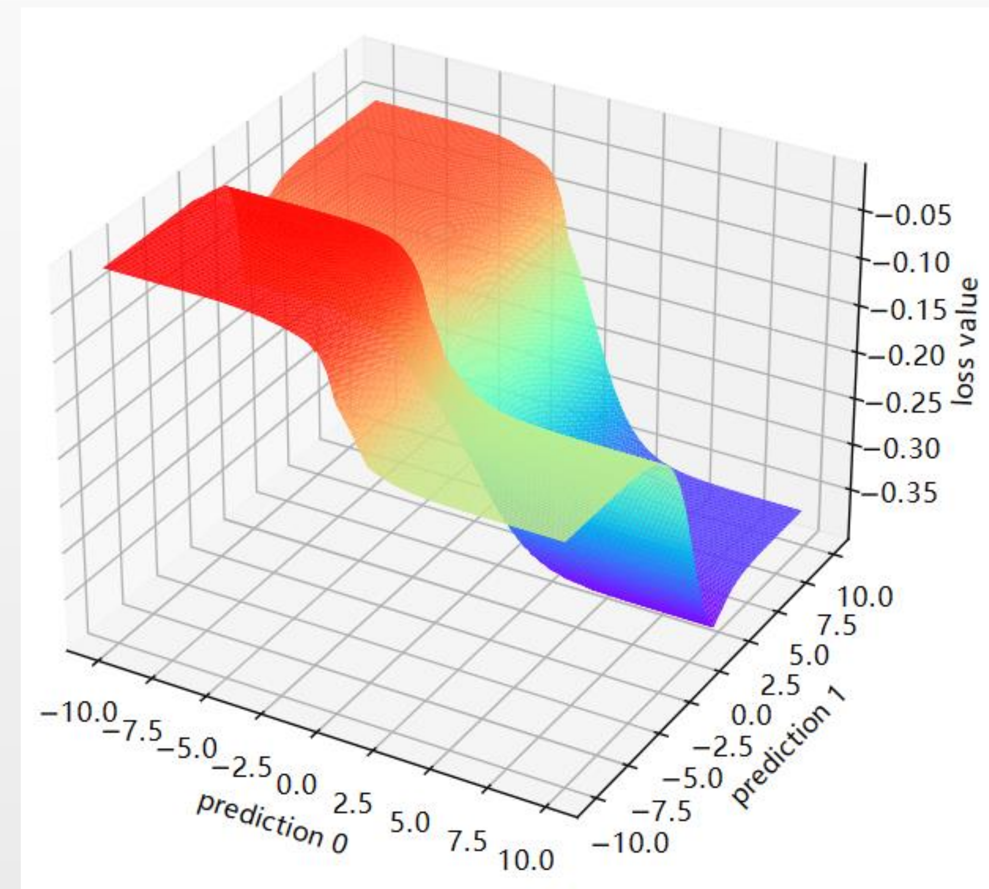
Visualization of DLN during MNIST learning



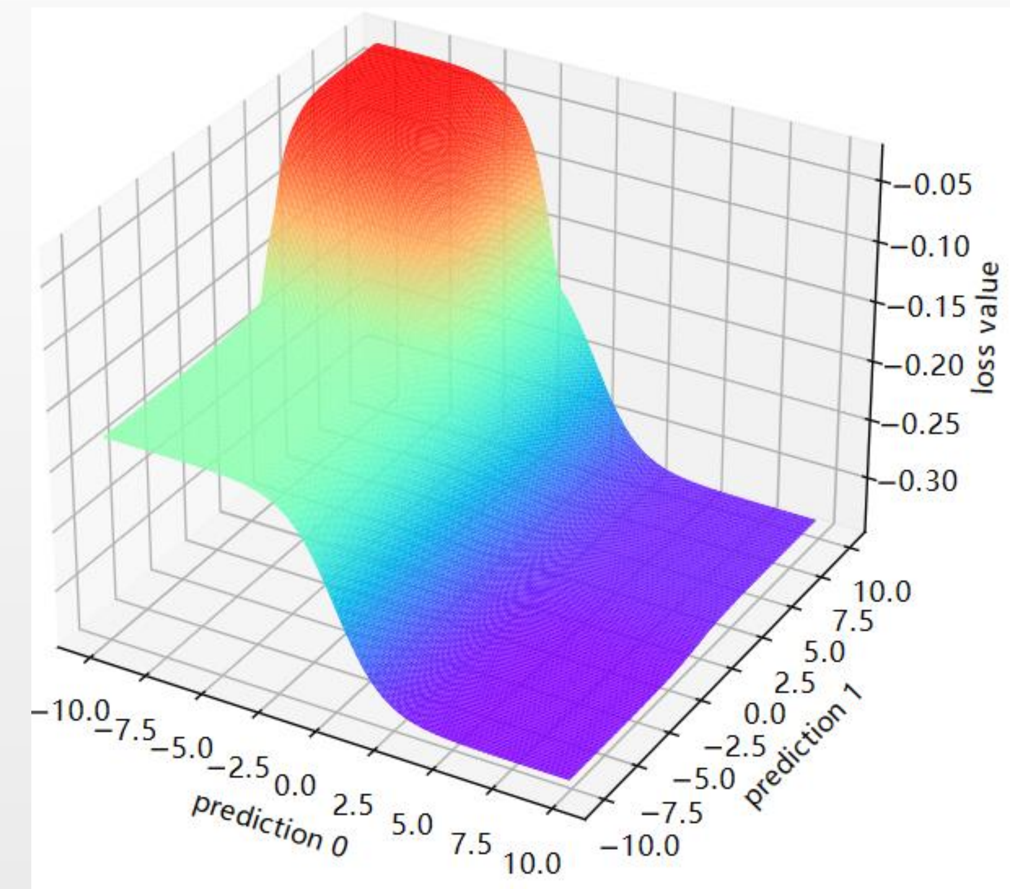
(a)



(b)

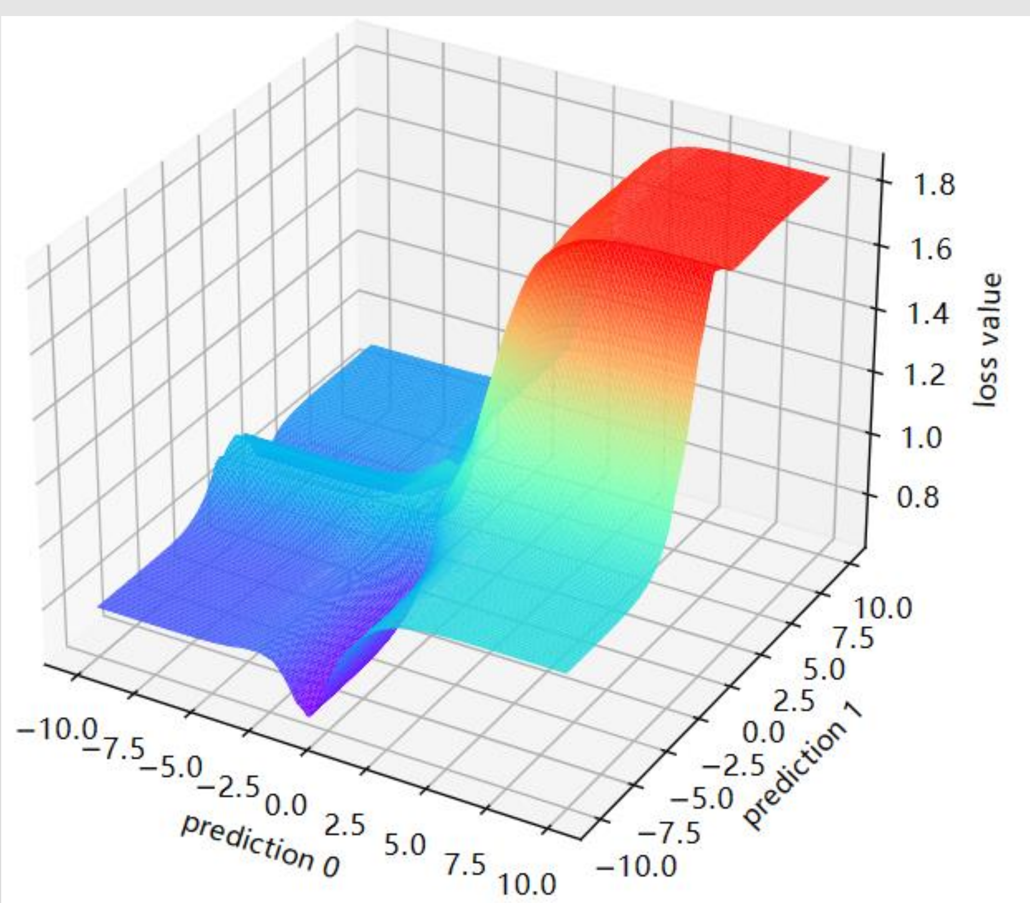


(c)

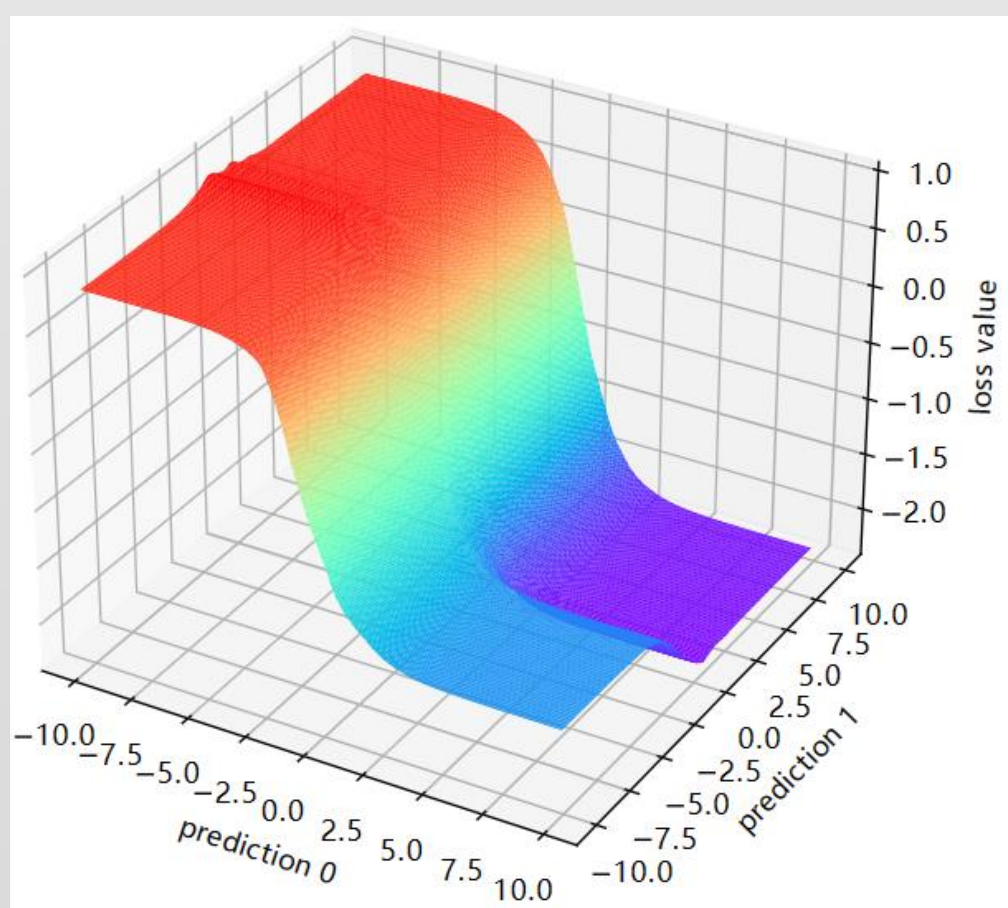


(d)

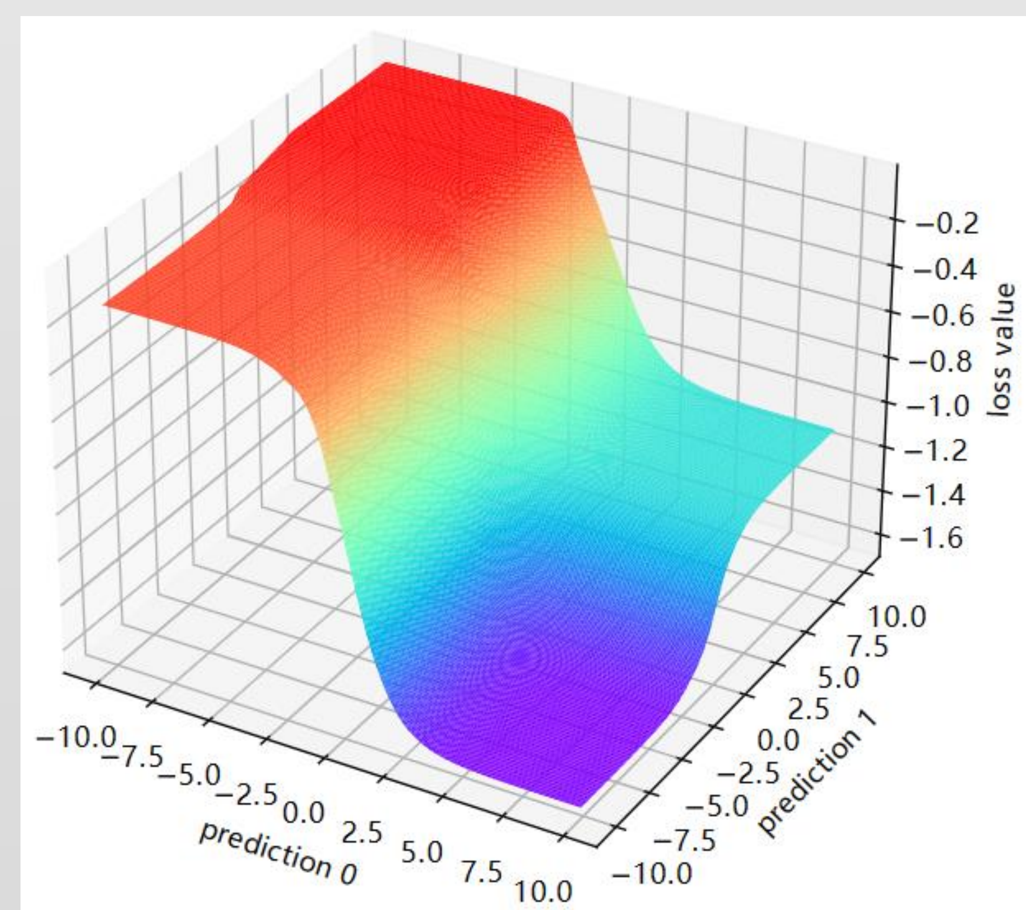
Visualization of DLN during CIFAR10 learning



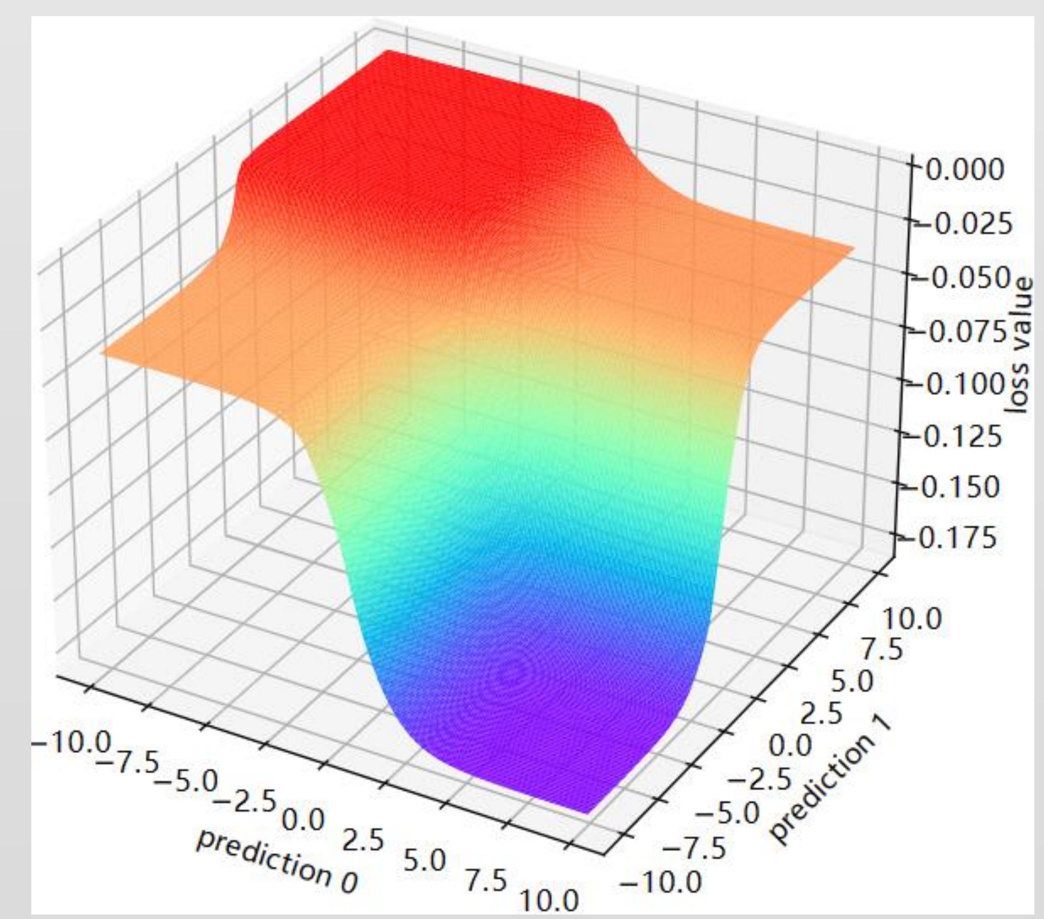
(a)



(b)



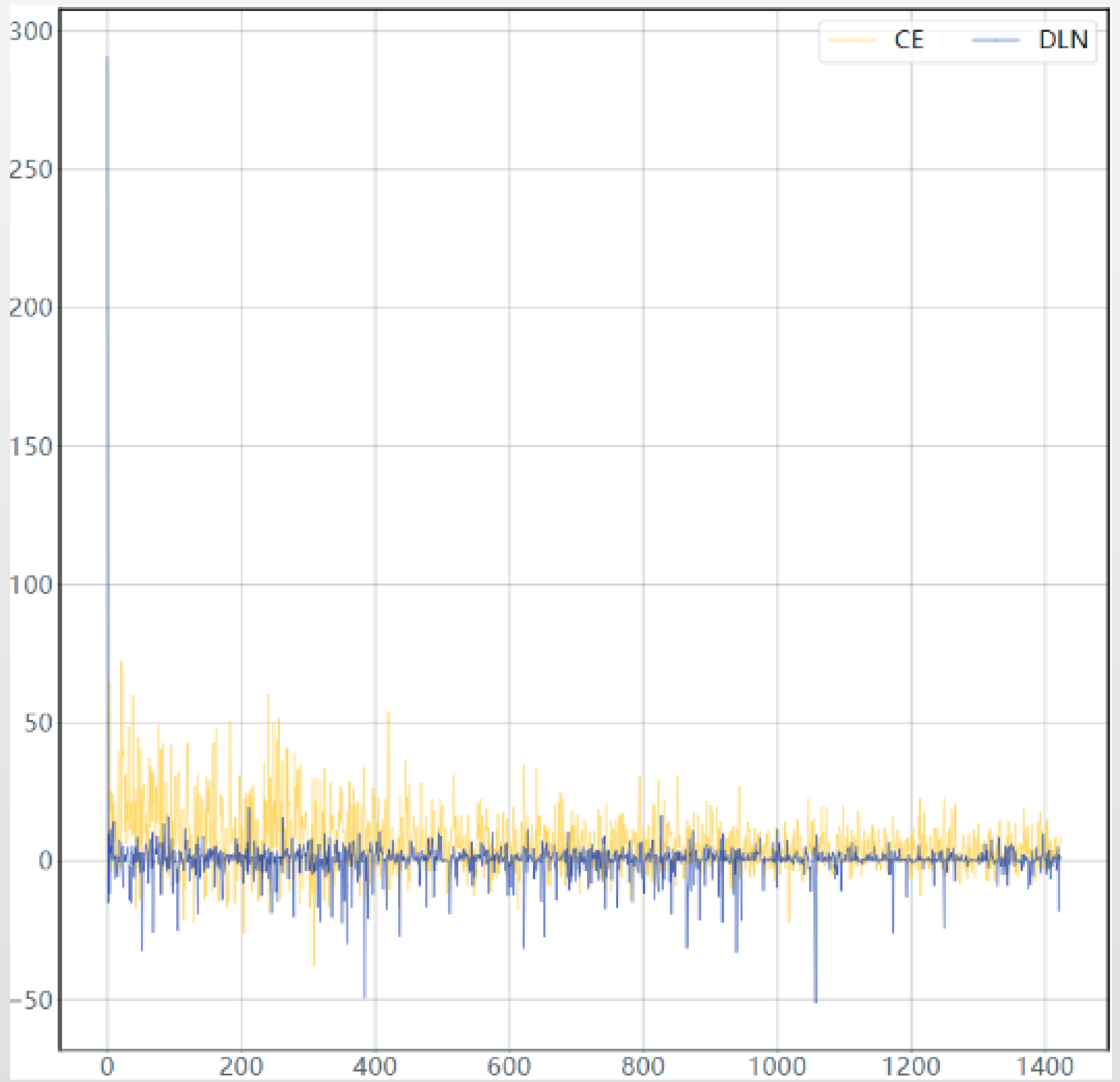
(c)



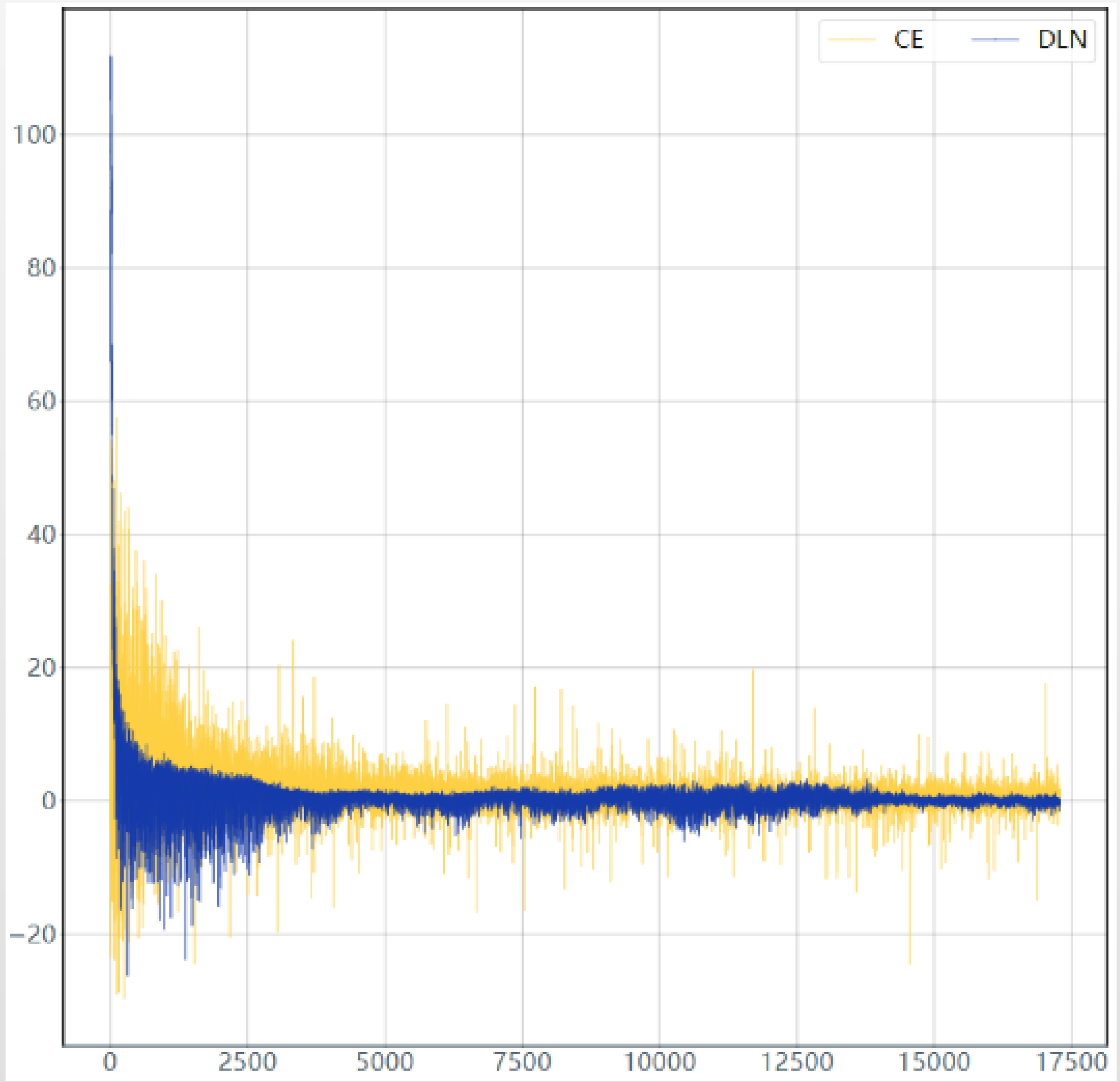
(d)

More details

Visualization of gradient of the student in noisy-label classification



(a) CIFAR10



(b) CIFAR100

References

[1] Jonathan T Barron. A general and adaptive robust loss function. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4331–4339, 2019.

[2] Yang Fan, Yingce Xia, Lijun Wu, Shufang Xie, Weiqing Liu, Jiang Bian, Tao Qin, and Xiang-Yang Li. Learning to reweight with deep interactions. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 7385–7393, 2021.

[3] Chen Huang, Shuangfei Zhai, Walter Talbott, Miguel Bautista Martin, Shih-Yu Sun, Carlos Guestrin, and Josh Susskind. Addressing the loss-metric mismatch with adaptive loss alignment. In International conference on machine learning, pages 2891–2900. PMLR, 2019.

[4] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In International conference on machine learning, pages 2304–2313. PMLR, 2018.

[5] Qingliang Liu and Jinmei Lai. Stochastic loss function. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 4884–4891, 2020.

[6] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In International Conference on Machine Learning, pages 507–516. PMLR, 2016.

[7] Tan Nguyen and Scott Sanner. Algorithms for direct 0–1 loss optimization in binary classification. In International Conference on Machine Learning, pages 1085–1093. PMLR, 2013.

[8] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[9] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. Advances in neural information processing systems, 32, 2019.

[10] Lijun Wu, Fei Tian, Yingce Xia, Yang Fan, Tao Qin, Lai Jian-Huang, and Tie-Yan Liu. Learning to teach with dynamic loss functions. Advances in neural information processing systems, 31, 2018.

[11] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2881–2890, 2017.

Thank you!