Meta-Weight-Net: Learning an Explicit Mapping For Sample Weighting

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, Deyu Meng*

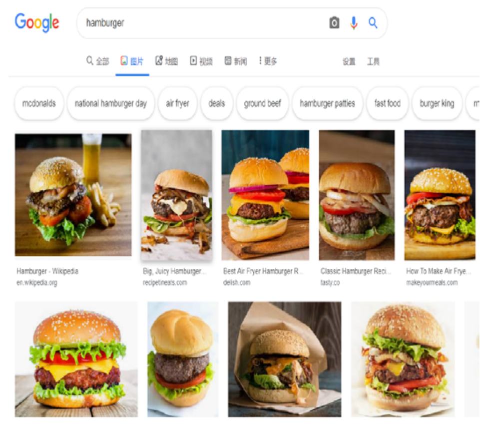
Introduction

Problem

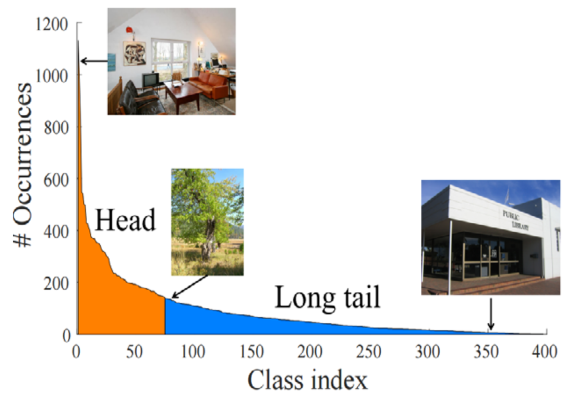
DNNs can easily overfit to **biased training data**.

- ✓ **Corrupted labels:** Data are collected from a crowd-sourcing system, web crawler, etc.
- ✓ **Class imbalance:** Real-world datasets are usually depicted as a long-tailed distribution.

Crowd-sourcing



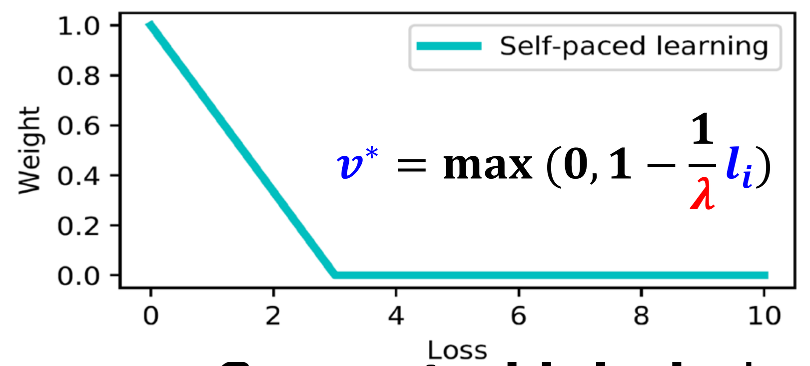
Web crawler



Long-tail distribution of SUN-397 dataset

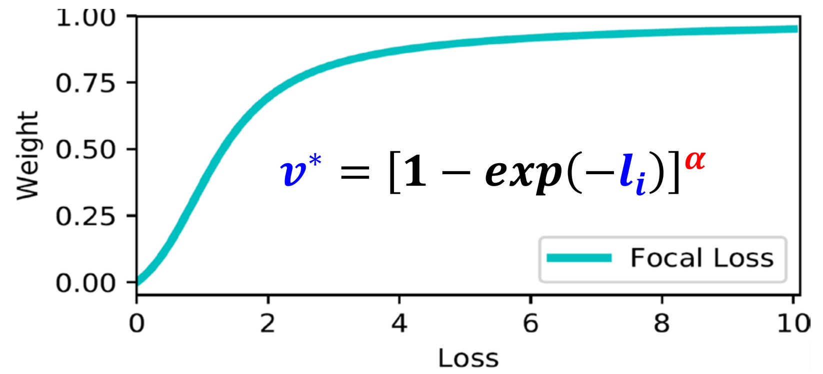
Motivation

- ✓ **Sample reweighting** is a commonly used strategy against this robust learning issue.
- ✓ There exist two entirely contradictive ideas for constructing the weighting function.



Corrupted labels ↓

E.g., SPL[1,2], iterative reweighting[3]



Class imbalance↑

E.g., AdaBoost[4,5], hard negative mining[6], focal loss[7].

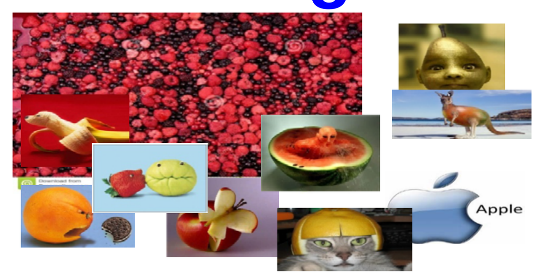
- ●**Need to pre-specify the form of weighting function based on certain assumptions on training data.**
- ●**Need to manually set hyper-parameters, raising their difficulty to be readily used in real applications.**

Meta-Weight-Net

The Meta-learning Objective

✓ We minimize the weighted training loss for classifier's updating.

Training Data



Meta Data

$$w^*(\Theta) = \arg\min_{w} \frac{1}{N} \sum_{i=1}^{N} \mathcal{V}(L_i^{\text{train}}(w); \Theta) L_i^{\text{train}}(w)$$

We formulate $\mathcal{V}(L_i^{train}(w); \Theta)$ as a MLP network called Meta-Weight-Net.

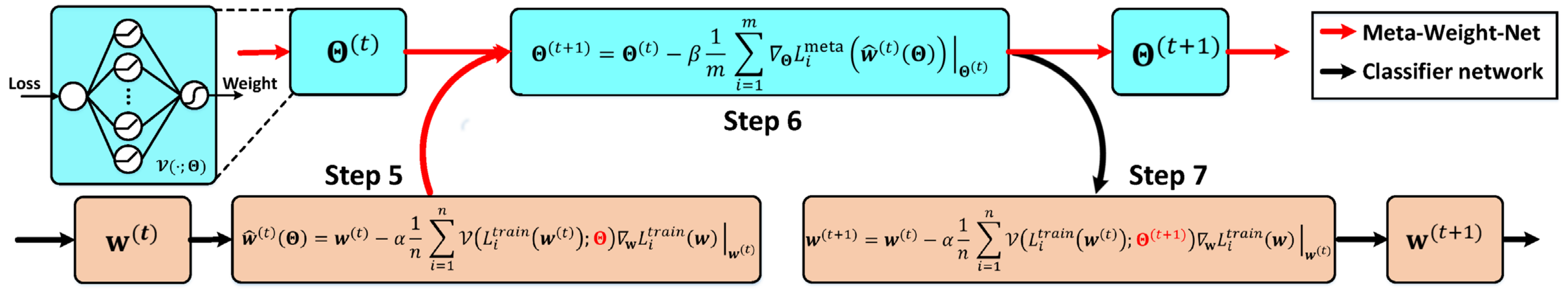
✓ The parameter Θ^* is obtained by minimizing the loss on meta data

$$\Theta^* = \arg\min_{\Theta} \frac{1}{M} \sum_{i=1}^{M} L_i^{\text{meta}}(w^*(\Theta))$$

References

1. Kumar, et al., Self-paced learning for latent variable models. In NeurIPS 2010.
2. Jiang et al,. Easy samples first. In ACM MM, 2014
3. Fernando et., A framework for robust subspace learning. IJCV, 2003
4. Freund et al., A decision-theoretic generalization of on-line learning and an application to boosting.1997.
5. Sun, et al., Cost-sensitive boosting for classification of imbalanced data. PR, 2007.
6. Malisiewicz et al.,. Ensemble of exemplar-svms for object detection and beyond. In ICCV, 2011.
7. Lin et al., Focal loss for dense object detection. In ICCV 2017.
8. Ren et al.,. Learning to reweight examples for robust deep learning. In ICML, 2018.
9. Jiang et al., Mentornet: Learning data-driven curriculum. In ICML, 2018.
10. Cui, et al., Class-balanced loss based on effective number of samples. In CVPR, 2019.

Algorithm



Code Downloading: https://github.com/xjtushujun/meta-weight-net

Please contact:
xjtushujun@gmail.com
dymeng@mail.xjtu.edu.cn

Convergence Analysis

Theorem 1. Suppose the loss function ℓ is Lipschitz smooth with constant L, and have ρ-bounded gradients with respect to training /meta data. $\mathcal{V}(\cdot)$ is differential with a δ-bounded gradient and twice differential with its Hessian bounded by $\mathcal{B}$. Let the learning rate α_t satisfies $\alpha_t = \min\{1, \frac{k}{T}\}$ for some $c > 0$, such that $\frac{\sigma\sqrt{T}}{c} \geq L$ and $\sum_{t=1}^{\infty} \beta_t \leq \infty, \sum_{t=1}^{\infty} \beta_t^2 \leq \infty$. Then the proposed algorithm can achieve $\mathbb{E}\left[\left\|\nabla \mathcal{L}^{\text{meta}}\left(\Theta^{(t)}\right)\right\|_2^2\right] \leq \epsilon$ in $\mathcal{O}(1/\epsilon^2)$ steps.

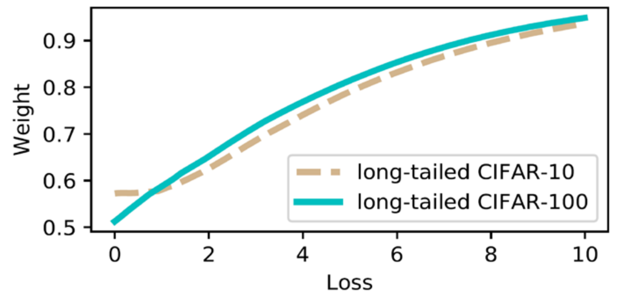
Theorem 2. The conditions in Theorem 1 hold, then we have: $\lim_{t\to\infty} \mathbb{E}\left[\left\|\nabla \mathcal{L}^{\text{meta}}\left(\mathrm{w}^{(t)}; \Theta^{(t+1)}\right)\right\|_2^2\right] = 0.$

Experimental Results

Class Imbalance Experiment

Test accuracy of ResNet-32 on long-tailed CIFAR-10 and CIFAR-100 [10]

Dataset Name	Long-Tailed CIFAR-10						Long-Tailed CIFAR-100					
Imbalance	200	100	50	20	10	1	200	100	50	20	10	1
BaseModel	65.68	70.36	74.81	82.23	86.39	92.89	34.84	38.32	43.85	51.14	55.71	70.50
Focal Loss	65.29	70.38	76.71	82.76	86.66	***93.03***	35.62	38.41	44.32	51.95	55.78	***70.52***
Class-Balanced	***68.89***	***74.57***	***79.27***	***84.36***	***87.49***	92.89	36.23	39.60	45.32	***52.59***	***57.99***	70.50
Fine-tuning	66.08	71.33	77.42	83.37	86.42	**93.23**	**38.22**	***41.83***	***46.40***	52.11	57.44	**70.72**
L2RW	66.51	74.16	78.93	82.12	85.19	89.25	33.38	40.23	44.44	51.64	53.73	64.11
Ours	**68.91**	**75.21**	**80.06**	**84.94**	**87.84**	92.66	***37.91***	**42.09**	**46.74**	**54.37**	**58.46**	70.37



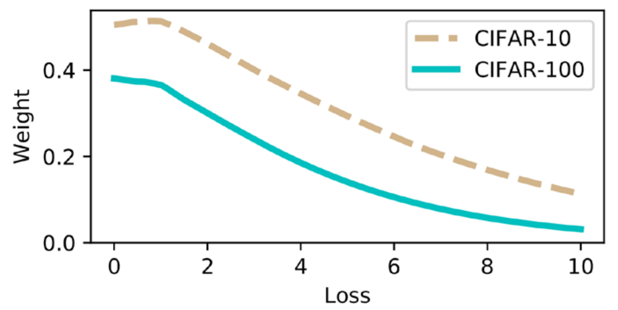
Corrupted Label Experiment

Test accuracy of WRN-28-10 with varying noise rates under uniform noise

Datasets	Noise Rate	BaseModel	Reed-Hard	S-Model	Self-paced	Focal Loss	Co-teaching	D2L	Fine-tining	MentorNet	L2RW	GLC	Ours
CIFAR-10	0%	95.60±0.22	94.38±0.14	83.79±0.11	90.81±0.34	**95.70±0.15**	88.67±0.25	94.64±0.33	***95.65±0.15***	94.35±0.42	92.38±0.10	94.30±0.19	94.52±0.25
	40%	68.07±1.23	81.26±0.51	79.58±0.33	86.41±0.29	75.96±1.31	74.81±0.34	85.60±0.13	80.47±0.25	87.33±0.22	86.92±0.19	***88.28±0.03***	**89.27±0.28**
	60%	53.12±3.03	73.53±1.54	—	53.10±1.78	51.87±1.19	73.06±0.25	68.02±0.41	78.75±2.40	82.80±1.35	82.24±0.36	***83.49±0.24***	**84.07±0.33**
CIFAR-100	0%	79.95±1.26	64.45±1.02	52.86±0.99	59.79±0.46	**81.04±0.24**	61.80±0.25	66.17±1.42	***80.88±0.21***	73.26±1.23	72.99±0.58	73.75±0.51	78.76±0.24
	40%	51.11±0.42	51.27±1.18	42.12±0.99	46.31±2.45	51.19±0.46	46.20±0.15	52.10 ±0.97	52.49±0.74	***61.39±3.99***	60.79±0.91	61.31±0.22	**67.73±0.26**
	60%	30.92±0.33	26.95±0.98	—	19.08±0.57	27.70±3.77	35.67±1.25	41.11±0.30	38.16±0.38	36.87±1.47	48.15±0.34	***50.81±1.00***	**58.75±0.11**

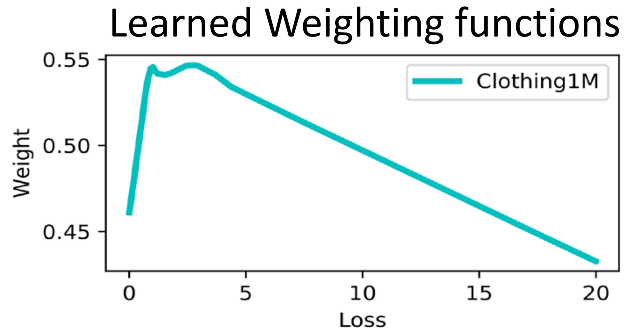
Test accuracy of ResNet-32 with varying noise rates under flip noise

Datasets	Noise Rate	BaseModel	Reed-Hard	S-Model	Self-paced	Focal Loss	Co-teaching	D2L	Fine-tining	MentorNet	L2RW	GLC	Ours
CIFAR-10	0%	***92.89±0.32***	92.31±0.25	83.61±0.13	88.52±0.21	93.03±0.16	89.87±0.10	92.02±0.14	**93.23±0.23**	92.13±0.30	89.25±0.37	91.02±0.20	92.04±0.15
	20%	76.83±2.30	88.28±0.36	79.25±0.30	87.03±0.34	86.45±0.19	82.83±0.85	87.66±0.40	82.47±3.64	86.36±0.31	87.86±0.36	***89.68±0.33***	**90.33±0.61**
	40%	70.77±2.31	81.06±0.76	75.73±0.32	81.63±0.52	80.45±0.97	75.41±0.21	83.89±0.46	74.07±1.56	81.76±0.28	85.66±0.51	**88.92±0.24**	***87.54±0.23***
CIFAR-100	0%	***70.50±0.12***	69.02±0.32	51.46±0.20	67.55±0.27	70.02±0.53	63.31±0.05	68.11±0.26	**70.72±0.22**	70.24±0.21	64.11±1.09	65.42±0.23	70.11±0.33
	20%	50.86±0.27	60.27±0.76	45.45±0.25	63.63±0.30	61.87±0.30	54.13±0.55	63.48±0.53	56.98±0.50	61.97±0.47	57.47±1.16	***63.07±0.53***	**64.22±0.28**
	40%	43.01±1.16	50.40±1.01	43.81±0.15	53.51±0.53	54.13±0.40	44.85±0.81	51.83±0.33	46.37±0.25	52.66±0.56	50.98±1.55	**62.22±0.62**	***58.64±0.47***



Clothing1M Experiment:

#	Method	Accuracy	#	Method	Accuracy
1	Cross Entropy	68.94	5	Joint Optimization [66]	72.23
2	Bootstrapping [58]	69.12	6	LCCN [67]	73.07
3	Forward [65]	69.84	7	MLNT [68]	73.47
4	S-adaptation [12]	70.36	8	Ours	73.72

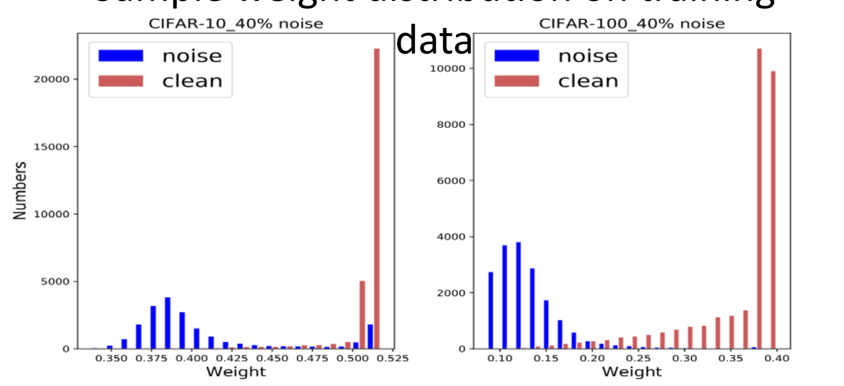


Ablation study

Test accuracy of different MW-Nets

architcture	Imbalance (factor 100)		Uniform noise (40%)		Flip noise (40%)	
	CIFAR10	CIFAR100	CIFAR10	CIFAR100	CIFAR10	CIFAR100
1-50-1	73.50	41.87	89.01	67.63	87.38	57.83
1-100-1	75.21	42.09	89.27	67.73	87.54	58.64
1-200-1	74.70	41.72	89.58	67.84	87.74	58.41
1-100-100-1	75.01	41.97	89.09	66.48	87.28	57.39
1-10-10-1	74.71	41.94	89.10	66.53	87.58	57.11
1-10-10-10-1	74.96	42.31	88.82	66.67	87.36	57.29

Sample weight distribution on training data



Weight variation curves on CIFAR10 dataset