JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

DISC: Learning from Noisy Labels via Dynamic Instance-Specific Selection and Correction

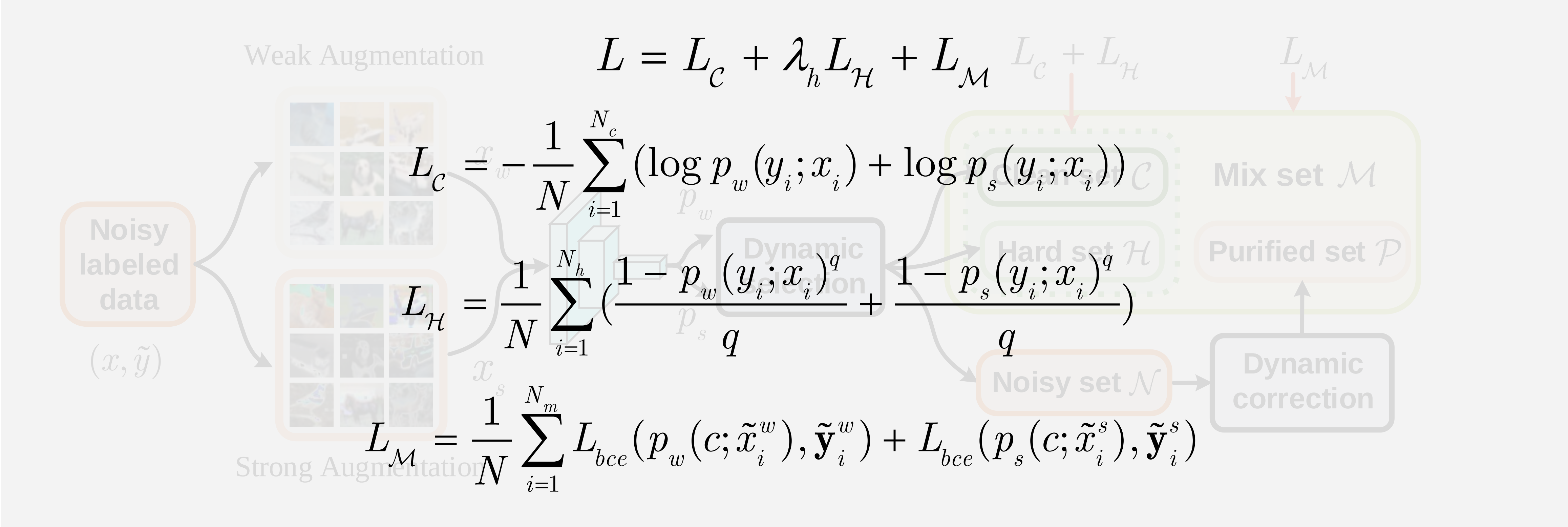
Yifan Li[1,2], Hu Han[1,2,3], Shiguang Shan[1,2,3], Xilin Chen[1,2]

1. Institute of Computing Technology, Chinese Academy of Sciences
2. University of Chinese Academy of Sciences
3. Peng Cheng Laboratory

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023

Overview

- Framework of **DISC**



$$L = L_{\mathcal{C}} + \lambda_h L_{\mathcal{H}} + L_{\mathcal{M}}$$

$$L_{\mathcal{C}} = -\frac{1}{N}\sum_{i=1}^{N_c}(\log p_w(y_i; x_i) + \log p_s(y_i; x_i))$$

$$L_{\mathcal{H}} = \frac{1}{N}\sum_{i=1}^{N_h}\left(\frac{1 - p_w(y_i; x_i)^q}{q} + \frac{1 - p_s(y_i; x_i)^q}{q}\right)$$

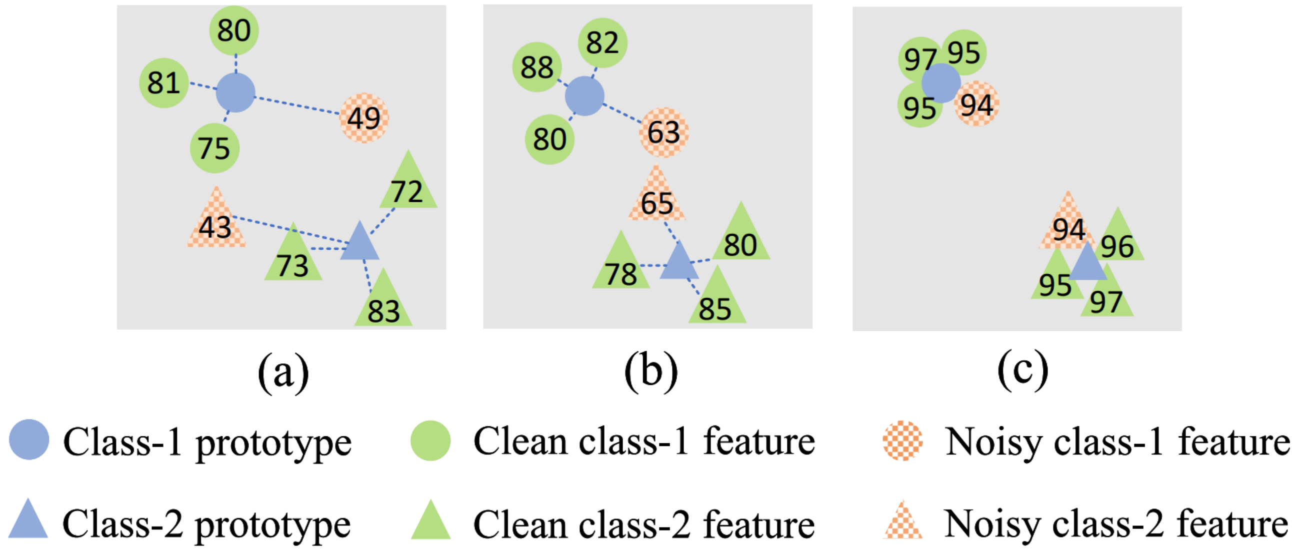
$$L_{\mathcal{M}} = \frac{1}{N}\sum_{i=1}^{N_m} L_{bce}(p_w(c; \tilde{x}_i^w), \tilde{\mathbf{y}}_i^w) + L_{bce}(p_s(c; \tilde{x}_i^s), \tilde{\mathbf{y}}_i^s)$$

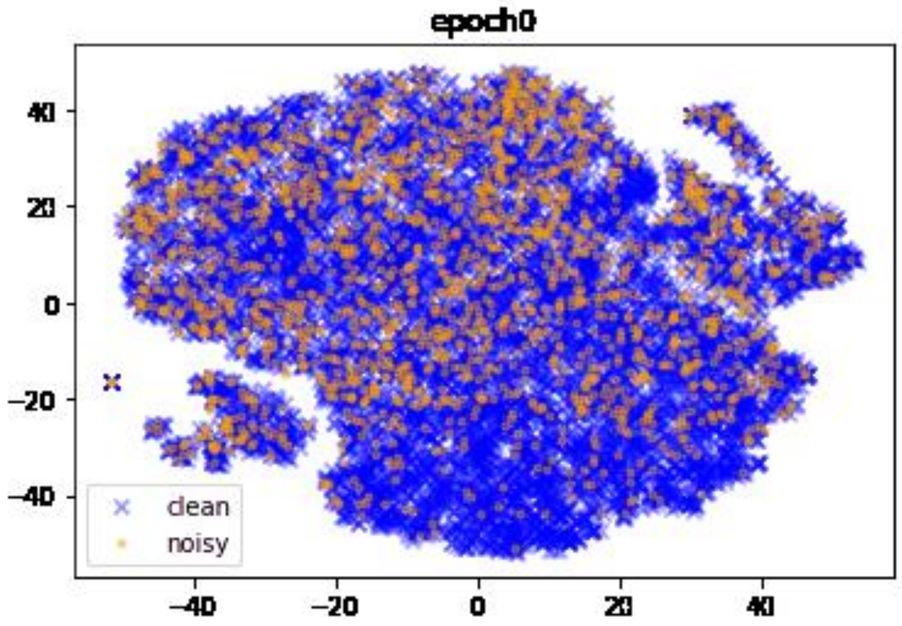
Contents

- Introduction
- Method
- Experiments
- Conclusion

Introduction

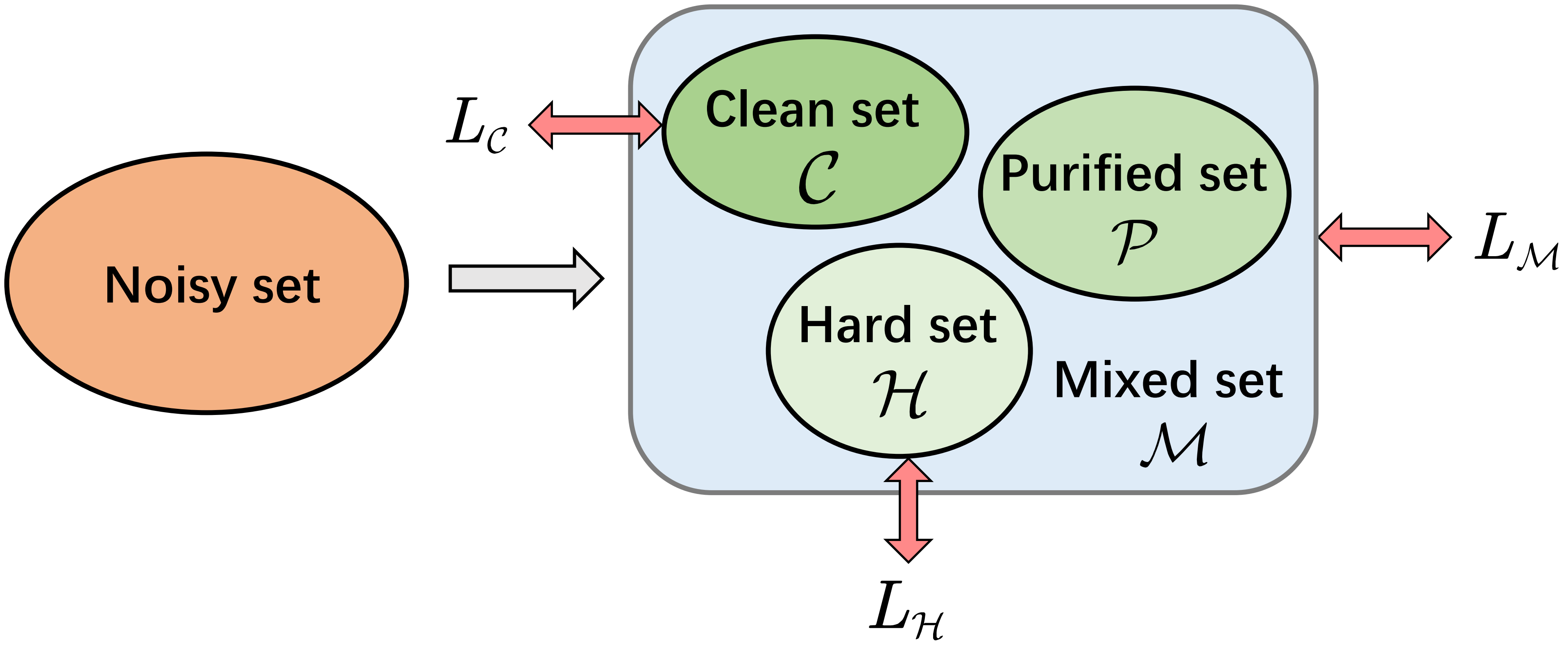
- Motivation
  - The memorization strength of DNNs towards different instances increases as training progresses





Introduction

- Motivation
  - Exploiting information in noisy set by "divide and conquer" strategy

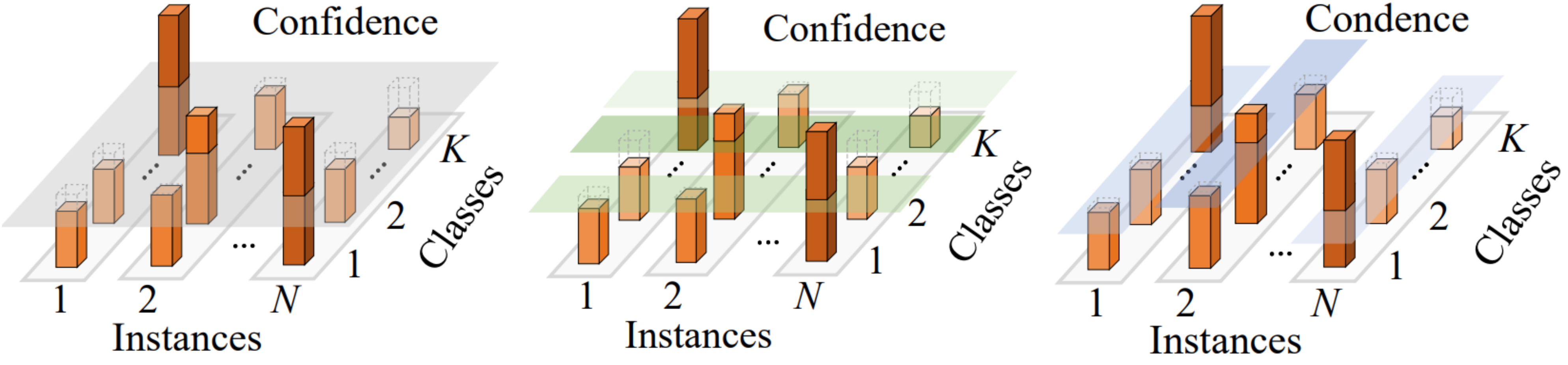


Introduction

- Contributions
  - The **memorization strength** of DNNs towards individual instances can be denoted by confidence, which increases along with training
  - **Dynamic instance-specific threshold** is proposed for selecting reliable labels and correcting noisy labels following an easy-to-hard curriculum
  - We propose a "**divide and conquer**" strategy. The dynamic threshold strategy is leveraged to group noisy data into three different subsets and different regularization strategies are utilized to handle individual subsets

Method

- Dynamic instance-specific threshold



(a) The global threshold (b) The Class-wise threshold (c) The dynamic instance-specific threshold

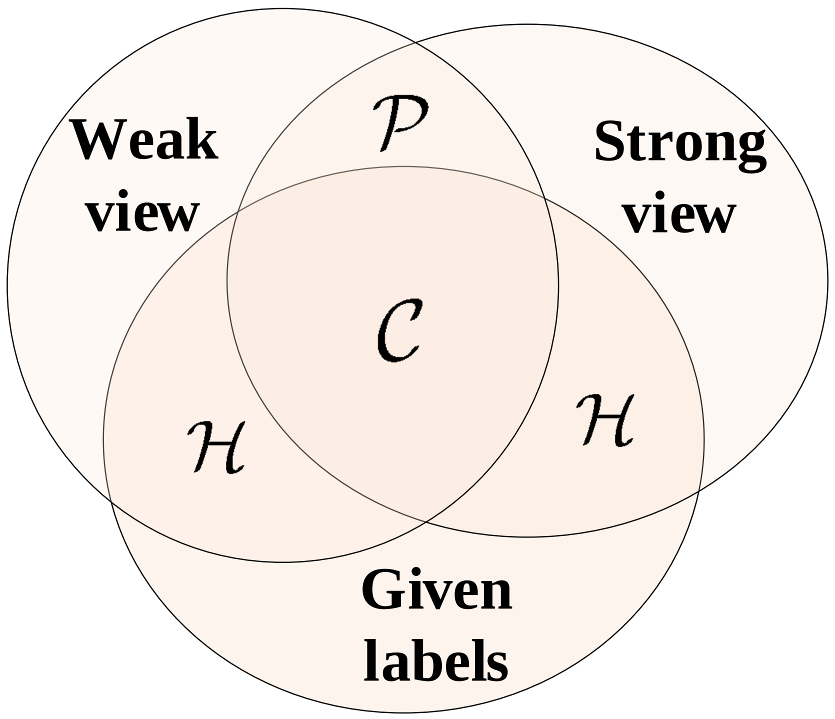
Momentum of confidence

$$\tau(t) = \lambda\tau(t-1) + (1-\lambda)p(t),\ \tau(0) = 0$$

$$\tau'(t) = max(\tau(t) + \sigma, 0.99)$$

Method

- Divide and conquer
  - Divide
    - The entire noisy set is divided into three different subsets according to the intersection of two views' predictions and the noisy labels



$$\mathcal{C} = \{x_i, y_i | p_w(y_i; x_i) > \tau_w(t)\} \cap \{x_i, y_i | p_s(y_i; x_i) > \tau_s(t)\}$$

$$\mathcal{H} = \{x_i, y_i | p_w(y_i; x_i) > \tau_w(t)\} \cup \{x_i, y_i | p_s(y_i; x_i) > \tau_s(t)\} - \mathcal{C}.$$

$$\mathcal{P} = \{x_i, \hat{y}_c = \arg\max_c p_{ws}(c; x_i) | \max_c p_{ws}(c; x_i) > \tau'(t), \forall c \in \mathcal{Y}\} - \{\mathcal{C} \cup \mathcal{H}\}$$

$$\mathcal{M} = \{\mathcal{C} \cup \mathcal{H} \cup \mathcal{P}\}$$

Method

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- Divide and conquer
  - Conquer
    - Different regularization strategies are adopted to conquer individual subsets

$$\mathcal{C} \longleftrightarrow L_{\mathcal{C}} = -\frac{1}{N}\sum_{i=1}^{N_c}(\log p_w(y_i;x_i) + \log p_s(y_i;x_i))$$

$$\mathcal{H} \longleftrightarrow L_{\mathcal{H}} = \frac{1}{N}\sum_{i=1}^{N_h}(\frac{1-p_w(y_i;x_i)^q}{q} + \frac{1-p_s(y_i;x_i)^q}{q})$$

$$\mathcal{M}(\mathcal{C}\cup\mathcal{H}\cup\mathcal{P}) \longleftrightarrow L_{\mathcal{M}} = \frac{1}{N}\sum_{i=1}^{N_m} L_{bce}(p_w(c;\tilde{x}_i^w),\tilde{\mathbf{y}}_i^w) + L_{bce}(p_s(c;\tilde{x}_i^s),\tilde{\mathbf{y}}_i^s)$$

Experiments

- Single noisy label image classification
  - Datasets

Datasets	# Class	Scale	Noise ratio	Noise sources
CIFAR10	10	60K	$\rho \in \{20\%, 40\%, 60\%\}$	Inst.
CIFAR100	100	60K	$\rho \in \{20\%, 40\%, 60\%\}$	Inst.
Tiny-ImageNet	200	120K	$\rho \in \{20\%, 50\%, 45\%\}$	Sym., asym.
Clothing1M	14	1,074K	38.5%	Real-world
WebVision	50	100K	20%	Real-world
Food101N	101	101K	18.4%	Real-world
Animals-10N	10	55K	8%	Real-world

  - Evaluation metric $$Acc = \frac{\#True}{\#Total}$$

Experiments

- Comparison with SOTA methods on CIFAR with Inst. noise

Dataset	CIFAR-10			CIFAR-100		
Noise type	Inst. 20%	Inst. 40%	Inst. 60%	Inst. 20%	Inst. 40%	Inst. 60%
CE*	83.93 ± 0.15	67.64 ± 0.26	43.83 ± 0.33	57.35 ± 0.08	43.17 ± 0.15	24.42 ± 0.16
Forward T [36]	87.22 ± 1.60	79.37 ± 2.72	66.56 ± 4.90	58.19 ± 1.37	42.80 ± 1.01	27.91 ± 3.35
DMI [36]	88.57 ± 0.60	82.82 ± 1.49	69.94 ± 1.34	57.90 ± 1.21	42.70 ± 0.92	26.96 ± 2.08
Mixup* [56]	87.71 ± 0.66	82.65 ± 0.38	58.59 ± 0.58	46.31 ± 0.25	45.14 ± 0.31	23.77 ± 0.26
GCE* [58]	89.80 ± 0.12	78.95 ± 0.15	60.76 ± 3.08	58.01 ± 0.26	45.69 ± 0.14	35.08 ± 0.23
Co-teaching [17]	88.87 ± 0.24	73.00 ± 1.24	62.51 ± 1.98	43.30 ± 0.39	23.21 ± 0.57	12.58 ± 0.58
Co-teaching+ [53]	89.80 ± 0.28	73.78 ± 1.39	59.22 ± 6.34	41.71 ± 0.78	24.45 ± 0.71	12.58 ± 0.58
JoCoR [47]	88.78 ± 0.15	71.64 ± 3.09	63.46 ± 1.58	43.66 ± 1.32	23.95 ± 0.44	13.16 ± 0.91
Reweight-R [49]	90.04 ± 0.46	84.11 ± 2.47	72.18 ± 2.47	58.00 ± 0.36	43.83 ± 8.42	36.07 ± 9.73
Peer Loss [33]	89.12 ± 0.76	83.26 ± 0.42	74.53 ± 1.22	61.16 ± 0.64	47.23 ± 1.23	31.71 ± 2.06
DivideMix [29]	93.33 ± 0.14	95.07 ± 0.11	85.50 ± 0.71	79.04 ± 0.21	76.08 ± 0.35	46.72 ± 1.32
CORSES^2 [7]	91.14 ± 0.46	83.67 ± 1.29	77.68 ± 2.24	66.47 ± 0.45	58.99 ± 1.49	38.55 ± 3.25
CAL [62]	92.01 ± 0.12	84.96 ± 1.25	79.82 ± 2.56	69.11 ± 0.46	63.17 ± 1.40	43.58 ± 3.30
CC [59]	93.68 ± 0.12	94.97 ± 0.09	94.95 ± 0.11	79.61 ± 0.19	76.58 ± 0.25	59.40 ± 0.46
DISC (ours)	$\mathbf{96.48 \pm 0.04}$	$\mathbf{95.94 \pm 0.04}$	$\mathbf{95.05 \pm 0.05}$	$\mathbf{80.12 \pm 0.13}$	$\mathbf{78.44 \pm 0.19}$	$\mathbf{69.57 \pm 0.14}$

Experiments

- Comparison with the SOTA methods on Tiny ImageNet with sym. and asym. noise.

Noise	Sym. 0%		Sym. 20%		Sym. 50%		Asym. 45%	
Methods	Avg.	Best	Avg.	Best	Avg.	Best	Avg.	Best
Standard	56.7	57.4	35.6	35.8	19.6	19.8	26.2	26.3
Decoupling [34]	-	-	36.3	37.0	22.6	22.8	26.1	26.6
F-Correction [36]	-	-	-	-	32.8	33.1	0.6	0.67
MentorNet [23]	-	-	-	-	35.5	35.8	26.2	26.6
Co-teaching+ [53]	52.1	52.4	47.7	48.2	41.2	41.8	26.5	26.9
M-Correction [1]	57.2	57.7	56.6	57.2	51.3	51.6	24.1	24.8
NCT [37]	61.5	62.4	57.2	58.2	47.4	47.8	42.4	43.0
UNICON [24]	62.7	63.1	58.4	59.2	52.4	52.7	-	-
DISC (Ours)	**68.2**	**68.5**	**67.5**	**67.9**	**63.9**	**64.3**	**52.8**	**53.6**

Experiments

- Comparison with the SOTA methods on Animals-10N, Food-101, WebVision and Clothing1M

Animals-10N

Method	Accuracy (%)
CE [11]	79.4 ± 0.14
GCE* (2018) [58]	81.5± 0.08
SELFIE (2019) [40]	81.8 ± 0.09
Mixup* (2017) [56]	82.7 ± 0.03
Co-learning (2021) [43]	83.0
PLC (2021) [57]	83.4 ± 0.43
Nested Co-teaching (2021) [6]	84.1 ± 0.1
GJS (2021) [11]	84.2 ± 0.07
DISC (ours)	**87.1 ± 0.15**

Food-101

Method	Accuracy (%)
CE [11]	81.67
CleanNet (2018) [28]	83.95
GCE* (2018) [58]	85.83
PLC (2021) [57]	83.4
GJS (2021) [11]	86.56
Mixup* (2017) [56]	87.34
Co-learning (2021) [43]	87.57
DISC (ours)	**89.02**

WebVision

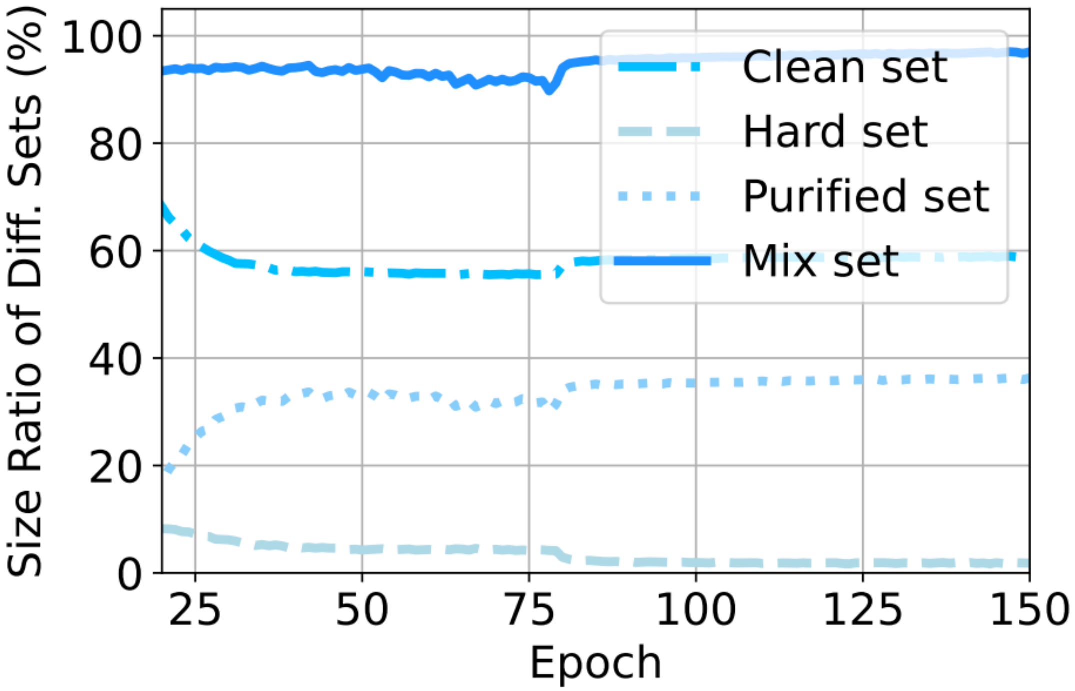
Dataset	WebVision		ILSVRC12	
Accuracy (%)	top1	top5	top1	top5
F-correction (2017) [36]	61.12	82.68	57.36	82.36
Decoupling (2017) [34]	62.54	84.74	58.26	82.26
D2L (2019) [27]	62.68	84.00	57.80	81.36
MentorNet [23]	63.00	81.40	57.80	79.92
Co-teaching (2018) [17]	63.58	85.20	61.48	84.70
INCV (2019) [5]	65.24	85.34	61.60	84.98
MentorMix (2020) [22]	76.0	90.2	72.9	91.1
ELR (2020) [32]	76.26	91.26	68.7	87.8
DivideMix (2020) [29]	77.32	91.64	75.20	90.84
ELR+ (2020) [32]	77.78	91.68	70.29	89.76
RRL (2021) [30]	77.8	91.3	74.4	90.9
GJS (2021) [11]	77.99	90.62	74.33	90.33
CC (2022) [59]	79.36	**93.64**	76.08	**93.86**
DISC (ours)	**80.28**	92.28	**77.44**	92.28

Clothing1M

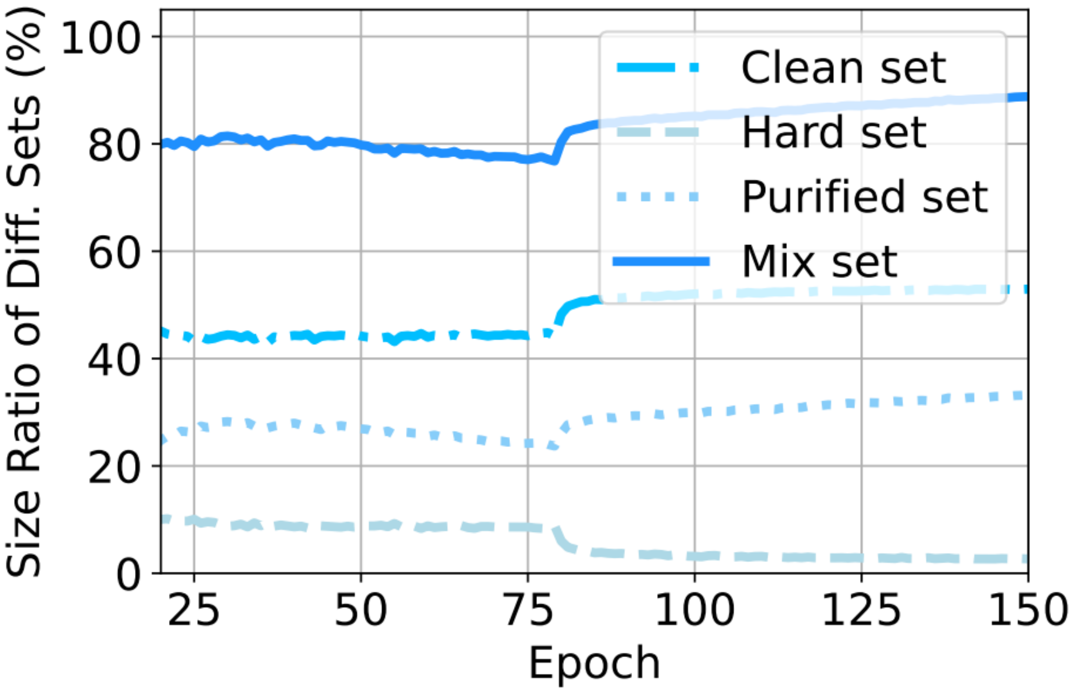
Method	Accuracy (%)
CE	68.94
Co-teaching (2018) [17]	69.21
JoCoR (2018) [47]	70.30
DMI (2019) [51]	72.46
DivideMix* (2019) [29]	74.45
ELR+* (2020) [32]	74.39
GJS (2021) [11]	71.64
CAL (2021) [62]	74.17
AugDesc* (2021) [35]	74.33
CC* (2022) [59]	74.54
DISC (ours)	73.72
DIST+DivideMix	**74.79**

Experiments

- The size ratio of different subsets on CIFAR (40% IDN)



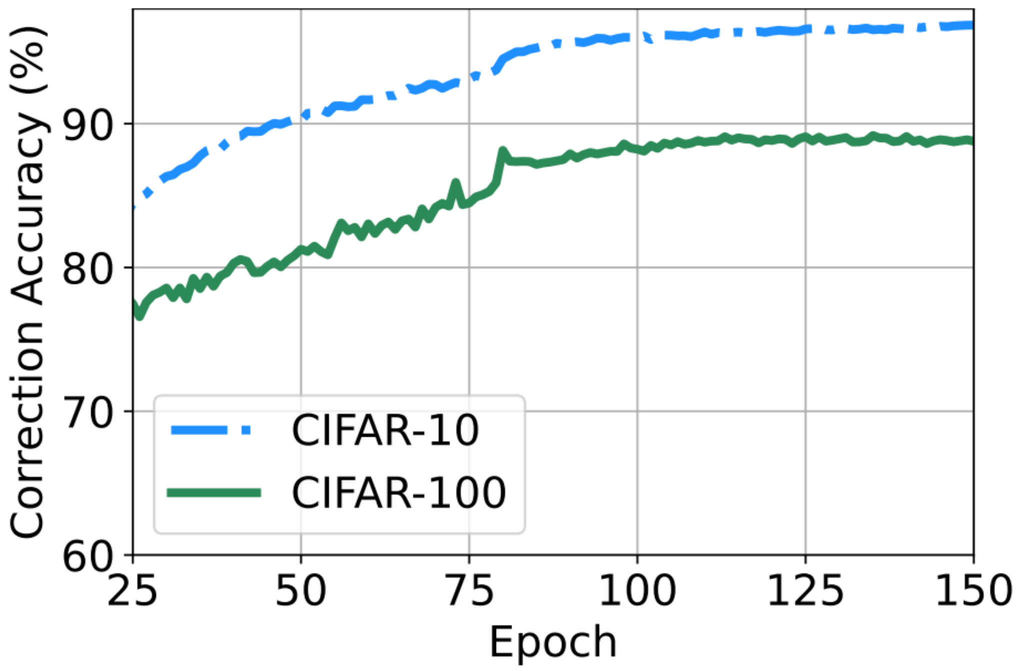
(a) CIFAR-10



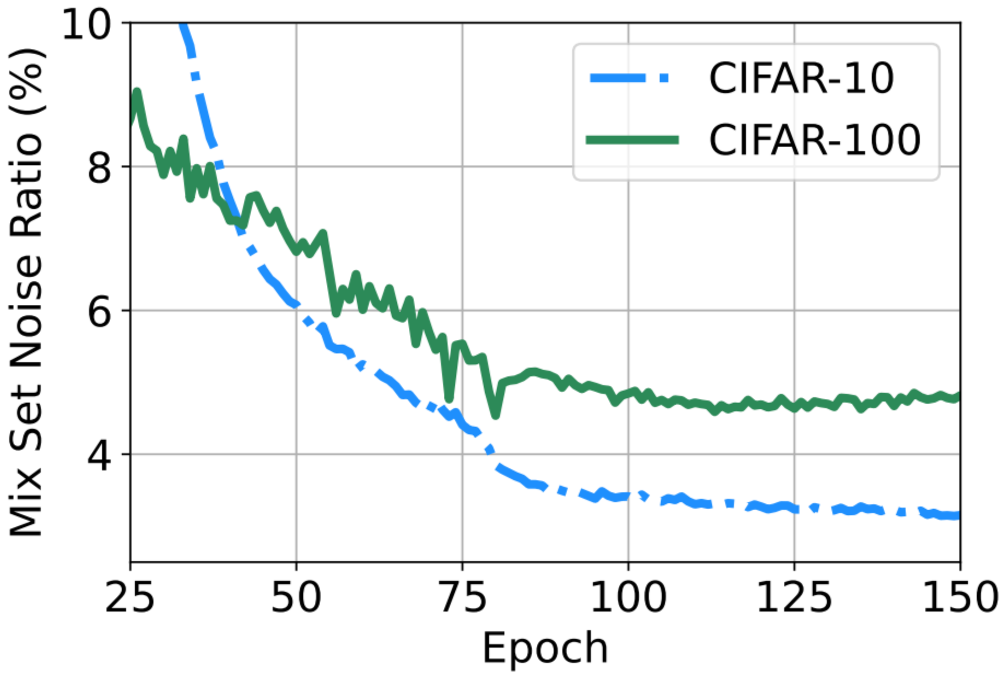
(b) CIFAR-100

Experiments

- The noise suppression on CIFAR (40% IDN)



(a) Correction acc. of $\mathcal{P}$



(b) Label noise rate in $\mathcal{M}$

Experiments

■ Ablation Study

☐ Ablation study on CIFAR under inst. noise 20%, 40% and 60%

Modules				CIFAR-10		CIFAR-100	
Two views	DIST	$\mathcal{H}$	$\mathcal{M}$	Inst. 20%	Inst. 40%	Inst. 20%	Inst. 40%
				83.93	67.63	53.35	43.16
✓				85.62	70.09	66.87	52.42
	✓			92.81	88.85	74.11	70.11
✓	✓			94.44	92.80	76.39	72.41
✓	✓	✓		94.52	92.82	76.45	72.51
✓	✓		✓	96.31	95.74	79.88	78.29
✓	✓	✓	✓	96.48	95.94	80.12	78.44

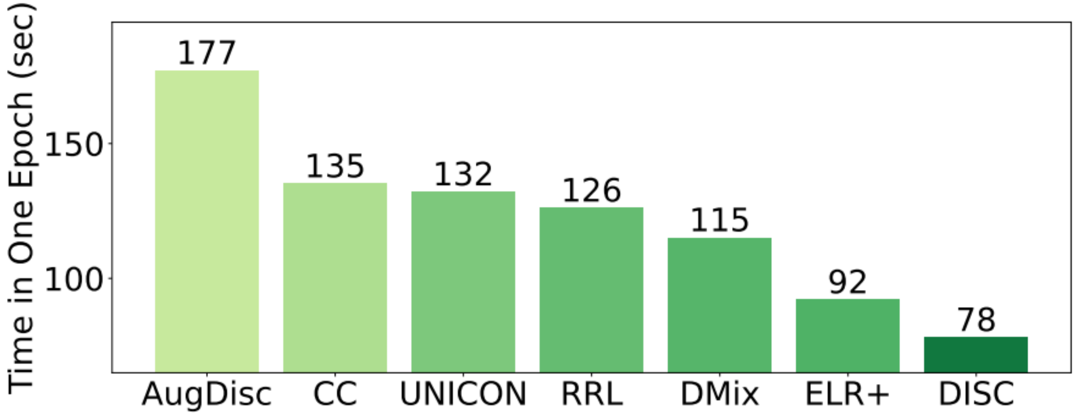
☐ Test acc. of the different views on CIFAR

Dataset	CIFAR-10			CIFAR-100		
Noise type	Inst. 20%	Inst. 40%	Inst. 60%	Inst. 20%	Inst. 40%	Inst. 60%
DISC-W	95.73	94.32	88.37	78.18	75.21	66.88
DISC-WW	96.20	94.47	93.65	79.24	77.67	68.85
DISC-WS	**96.48**	**95.94**	**94.86**	**80.12**	**78.44**	**69.57**

☐ Test acc. of different selection methods on CIFAR

Dataset	CIFAR-10			CIFAR-100		
Noise type	Inst. 20%	Inst. 40%	Inst. 60%	Inst. 20%	Inst. 40%	Inst. 60%
Small-losses [16]	90.83	84.81	21.47	71.82	63.89	22.56
GMM [28]	92.78	85.12	48.81	72.91	30.73	11.19
Fixed thres. 0.5 [29]	84.25	60.53	20.85	61.37	45.40	14.78
DIST	**92.81**	**88.85**	**80.66**	**74.11**	**70.01**	**60.07**

☐ Training and testing time profiling with PresNet-34 backbone and RTX 3090 GPU on CIFAR-10 with 20% inst. noise in one epoch



Conclusion

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- Memorization strength of DNNs towards individual instances could be reflected by confidences, which become higher along with training
- DISC is able to set a reasonable threshold for each instance and delicately divide the noisy data into different subsets, which can effectively suppress the label noise during classification learning
- However, DISC may also induce confirmation bias, since high-confidence instances may be the easy ones with noisy labels rather than the clean ones

Thank you for listening :)

Our paper and code are available:

Paper

Code

Feel free to contact **Yifan Li** via:

liyifan20g@ict.ac.cn

Introduction

- Backgrounds
  - Label noise widely exists in the test sets of different datasets

Dataset	Modality	Size	Model	Test Set Errors: CL guessed	MTurk checked	validated	estimated	% error
MNIST	image	10,000	2-conv CNN	100	100 (100%)	15	-	0.15
CIFAR-10	image	10,000	VGG	275	275 (100%)	54	-	0.54
CIFAR-100	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-	5.85
Caltech-256†	image	29,780	Wide ResNet-50-2	2,360	2,360 (100%)	458	-	1.54
ImageNet*	image	50,000	ResNet-50	5,440	5,440 (100%)	2,916	-	5.83
QuickDraw†	image	50,426,266	VGG	6,825,383	2,500 (0.04%)	1870	5,105,386	10.12
20news	text	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-	1.09
IMDB	text	25,000	FastText	1,310	1,310 (100%)	725	-	2.90
Amazon Reviews†	text	9,996,437	FastText	533,249	1,000 (0.2%)	732	390,338	3.90
AudioSet	audio	20,371	VGG	307	307 (100%)	275	-	1.35

*Because the ImageNet test set labels are not publicly available, the ILSVRC 2012 validation set is used.

†Because no explicit test set is provided, we study the entire dataset to ensure coverage of any train/test split.

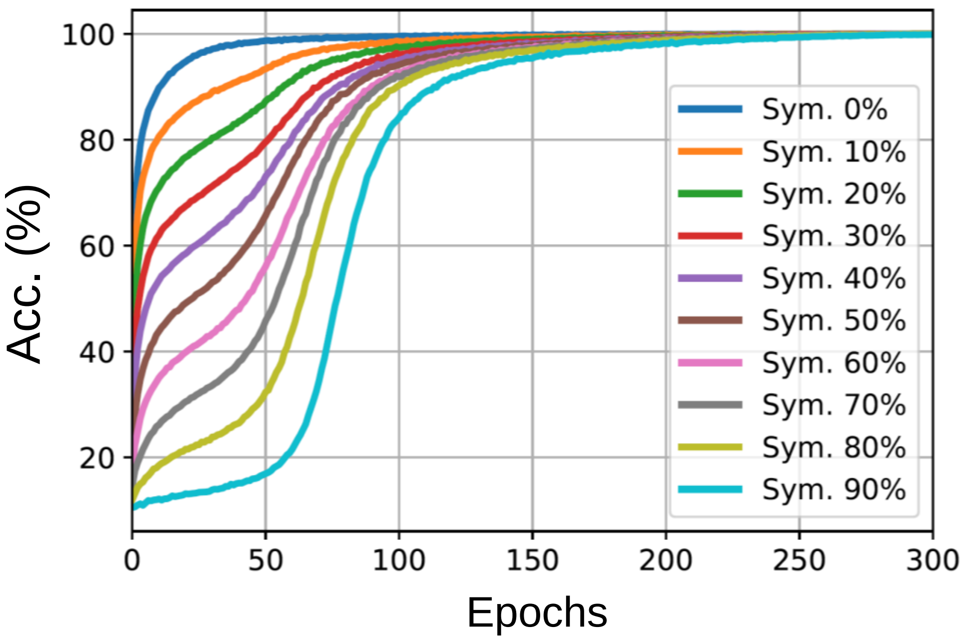
There also exists label noise in the validation set

The training set may be even noisier than the test set !

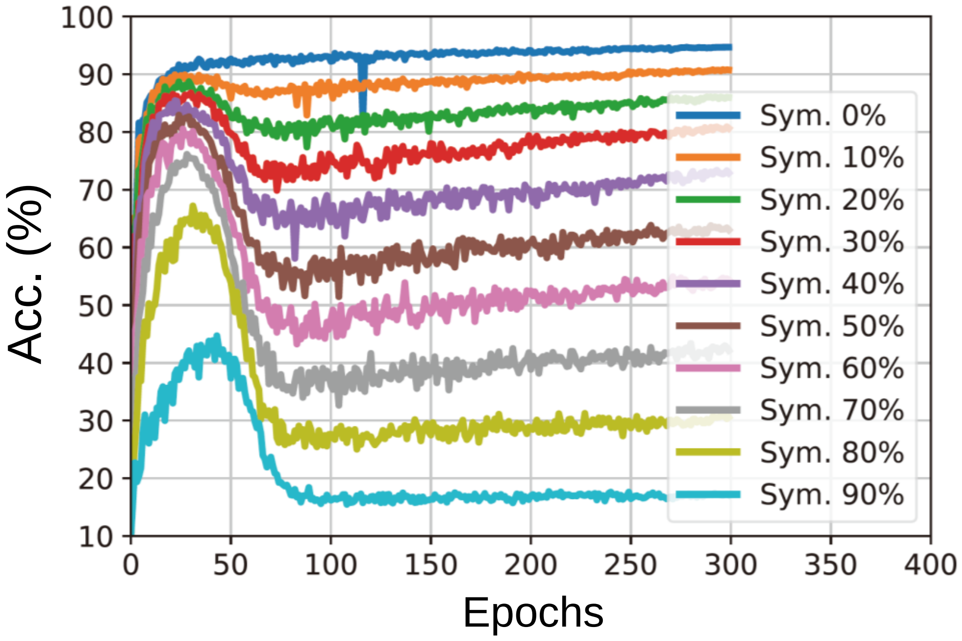
Northcutt, Curtis G., Anish Athalye, and Jonas Mueller. "Pervasive label errors in test sets destabilize machine learning benchmarks." *arXiv preprint arXiv:2103.14749* (2021).

Introduction

- Backgrounds
  - Label noise will harm the generalization ability of model
    - The model selected by validation set is sub-optimal
    - DNNs tend to **memorize** the label noise in the training set



(a) Training set accuracy



(a) Test set accuracy