ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

Understanding and Improving Early Stopping for Learning with Noisy Labels

NeurIPS 2021

Motivation

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

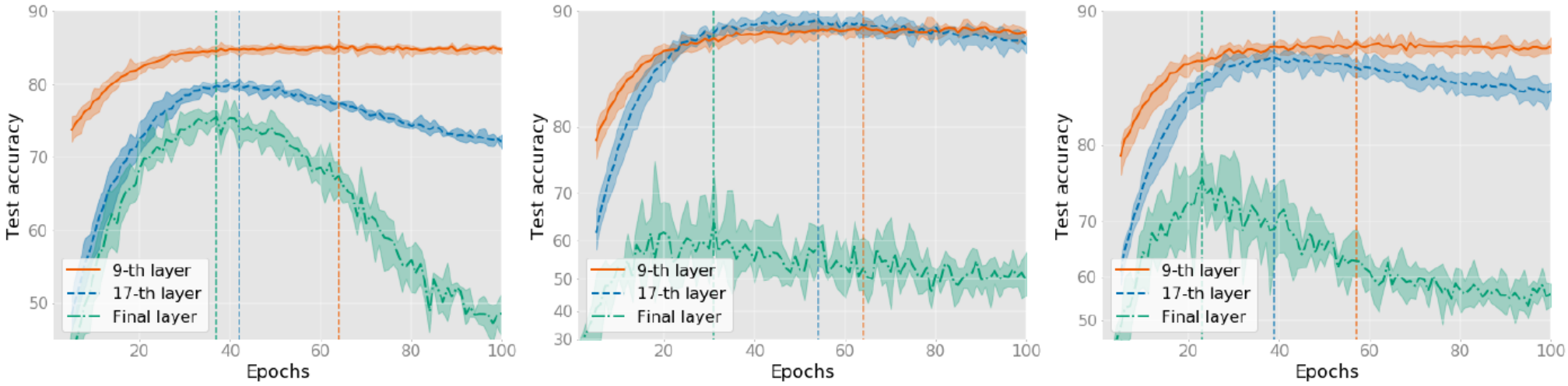
- Memorization effect of DNNs

  deep neural networks tend to first memorize and fit easy (clean) examples and then overfit hard (noisy) examples.

- To exploit the memorization effect, a core issue is to study when to stop the optimization of the network.
- Current methods usually adopt an early stopping strategy, which decides the stopping point by considering the network as a whole.

- DNNs trained By SGD, supervisory signals will gradually propagate through the whole network from latter layers to former layers.

Observation



(a) Symmetric 50% (b) Pairflip 45% (c) Instance 40%

Figure 1: We train a ResNet-18 model on CIFAR-10 with three types of noisy labels and evaluate the impact of noisy labels on the representations from the 9-th layer, the 17-th layer, and the final layer. The X-axis is the number of epochs for the first block of the network. The curves present the mean of five runs and the best performances are indicated with dotted vertical lines.

Method

- Objective function

$$\min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(\boldsymbol{x}_i; \Theta), \tilde{y}_i),$$

- Training model for a relatively small epoch number T.

- The network can be constituted with L DNN parts

$$\begin{aligned} \boldsymbol{z}_1 &= f_1(\boldsymbol{x}; \Theta_1), \\ \boldsymbol{z}_l &= f_l(\boldsymbol{z}_{l-1}; \Theta_l), \quad l = 2, \ldots, L \end{aligned}$$

Method

- For the first part, train T_1 epochs with the following objective.

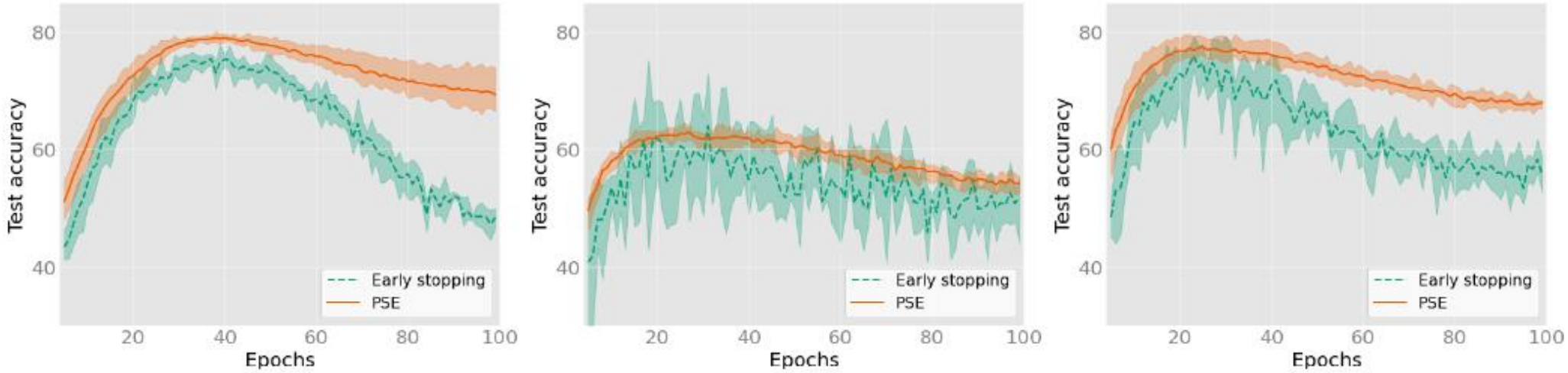
$$\min_{\Theta_1 \ldots \Theta_k} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(\boldsymbol{x}_i; \Theta_1, \ldots, \Theta_L), \tilde{y}_i).$$

- Then keep the obtained parameter fixed; optimize the $l\text{-}th$ DNN part with T_l epochs.

$$\min_{\Theta_l \ldots \Theta_k} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(\boldsymbol{x}_i; \Theta_1^*, \ldots, \Theta_{l-1}^*, \Theta_l, \ldots, \Theta_L), \tilde{y}_i), \quad l = 2 \ldots L$$

$$T_1 \geq T_2 \geq \cdots \geq T_L$$

Compared with traditional early stopping



(a) Symmetric 50% (b) Pairflip 45% (c) Instance 40%

Figure 2: Performance of the traditional early stopping trick and the proposed PES on CIFAR-10 with different types of label noise. The lines present the mean of five runs.

Learning with Confident Examples

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

- Select confident examples to facilitate the model training.
- To make the results more robust by generating two different augmentations.

$$\mathcal{D}_l = \{(\boldsymbol{x}_i, \tilde{y}_i) | \tilde{y}_i = \hat{y}_i, i = 1, \ldots, n\},$$

$$\hat{y}_i = \underset{k \in \{1,\ldots,K\}}{\arg\max} \frac{1}{2}[f^k(\text{Augment}(\boldsymbol{x}_i); \Theta) + f^k(\text{Augment}(\boldsymbol{x}_i); \Theta)],$$

- Train the DNN based on confident examples set with the following objective.

$$\mathcal{L}_c = \sum_{i=1}^{N} w_{y_i} \mathcal{L}_p(\tilde{y}_i, f(\boldsymbol{x}_i; \Theta)), \tag{6}$$

where w_i is the corresponding class weight. Assuming that $\sigma_k = |\{(\boldsymbol{x}_i, \tilde{y}_i) | \tilde{y}_i = k, (\boldsymbol{x}_i, \tilde{y}_i) \in \mathcal{D}_l\}|$ denotes the cardinality of the confident example set belonging to the k-th class. Then, we can set $w_i = \sigma_i / (\sum_{j=1}^{K} \sigma_j)$ to indicate the corresponding class importance.

Combining with Semi-Supervised Learning

- Training with only confident examples neglects the rest data may suffer from insufficient training examples.

- Using semi-supervised techniques (MixMatch) by regarding the noisy examples as unlabeled data.

$$\begin{cases} \mathcal{D}_l = \{(\boldsymbol{x}_i, \tilde{y}_i) | \tilde{y}_i = \hat{y}_i, i = 1, ...n\} \\ \mathcal{D}_u = \{\boldsymbol{x}_i | \tilde{y}_i \neq \hat{y}_i, i = 1, ...n\} \end{cases}$$

$$\hat{y}_i = \underset{k \in \{1,...K\}}{\arg\max} \frac{1}{2}[f^k(\text{Augment}(\boldsymbol{x}_i); \Theta) + f^k(\text{Augment}(\boldsymbol{x}_i); \Theta)], \tag{7}$$

Algorithm

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Algorithm 1: Progressive Early Stopping with Semi-Supervised Learning

Input: Neural network with trainable parameters $\Theta = \{\Theta_1, \ldots, \Theta_L\}$, Noisy training dataset $\{\boldsymbol{x}_i, \tilde{y}_i)\}_{i=1}^n$, Number of training epochs for different part: $T_1, \ldots, T_L$, and training epochs T_c for refining with confident examples.

for $i = 1, \ldots, T_1$ **do**
    Optimize network parameter Θ with Eq. (3);

for $l = 2, \ldots, L$ **do**
    Froze $\{\Theta_1, \ldots, \Theta_{l-1}\}$ and re-initialize $\{\Theta_l, \ldots, \Theta_L\}$;
    for $i = 1, \ldots, T_l$ **do**
        Optimize network parameter $\{\Theta_l, \ldots, \Theta_L\}$ with Eq. (4);

Unfroze Θ;

for $i = 1, \ldots, T_c$ **do**
    Extract confident example set $\mathcal{D}_l$ and unlabeled set $\mathcal{D}_u$ with classifier $f(\cdot, \Theta)$ by Eq. (7);
    Training the classifier $f(\cdot, \Theta)$ with MixMatch loss on $\mathcal{D}_l$ and $\mathcal{D}_u$;

Evaluate the obtained classifier $f(\cdot, \Theta)$.

Preliminary Experiments

Table 1: Preliminary analysis of the performance and the quality of extracted confident examples on CIFAR-10. The mean and standard deviation are computed over five runs.

Metrics	Methods	Sym-20%	Sym-50%	Pair-45%	Inst-20%	Inst-40%
Test Accuracy	Early Stopping	82.55±2.46	70.76±1.24	60.62±5.59	84.41±0.90	74.73±2.65
	PES	**85.87±1.59**	**75.87±1.33**	**62.40±2.34**	**86.58±0.45**	**77.07±1.18**
Label Precision	Early Stopping	98.81±0.15	94.65±0.19	72.53±5.26	**98.70±0.43**	**90.77±1.87**
	PES	**98.96±0.09**	**95.46±0.14**	**72.99±2.27**	98.52±0.19	90.63±0.92
Label Recall	Early Stopping	88.51±2.26	75.18±1.00	67.84±5.06	90.37±1.01	82.15±3.17
	PES	**92.67±1.43**	**81.03±1.83**	**71.06±2.27**	**93.24±0.60**	**85.91±0.68**

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Table 2: Comparison with state-of-the-art methods without semi-supervised learning on CIFAR-10 and CIFAR-100. The mean and standard deviation computed over five runs are presented.

Dataset	Method	Symmetric		Pairflip	Instance	
		20%	50%	45%	20%	40%
CIFAR10	CE	84.00±0.66	75.51±1.24	63.34±6.03	85.10±0.68	77.00±2.17
	Co-teaching	87.16±0.11	72.80±0.45	70.11±1.16	86.54±0.11	80.98±0.39
	Forward	85.63±0.52	77.92±0.66	60.15±1.97	85.29±0.38	74.72±3.24
	Joint Optim	89.70±0.11	85.00±0.17	82.63±1.38	89.69±0.42	82.62±0.57
	T-revision	89.63±0.13	83.40±0.65	77.06±6.47	90.46±0.13	85.37±3.36
	DMI	88.18±0.36	78.28±0.48	57.60±14.56	89.14±0.36	84.78±1.97
	CDR	89.72±0.38	82.64±0.89	73.67±0.54	90.41±0.34	83.07±1.33
	Ours	**92.38±0.40**	**87.45±0.35**	**88.43±1.08**	**92.69±0.44**	**89.73±0.51**
CIFAR100	CE	51.43±0.58	37.69±3.45	34.10±2.04	52.19±1.42	42.26±1.29
	Co-teaching	59.28±0.47	41.37±0.08	33.22±0.48	57.24±0.69	45.69±0.99
	Forward	57.75±0.37	44.66±1.01	27.88±0.80	58.76±0.66	44.50±0.72
	Joint Optim	64.55±0.38	50.22±0.41	42.61±0.61	65.15±0.31	55.57±0.41
	T-revision	65.40±1.07	50.24±1.45	41.10±1.95	60.71±0.73	51.54±0.91
	DMI	58.73±0.70	44.25±1.14	26.90±0.45	58.05±0.20	47.36±0.68
	CDR	66.52±0.24	55.30±0.96	43.87±1.35	67.33±0.67	55.94±0.56
	Ours	**68.89±0.45**	**58.90±2.72**	**57.18±1.44**	**70.49±0.79**	**65.68±1.41**

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Table 3: Comparison with state-of-the-art methods with semi-supervised learning on CIFAR-10 and CIFAR-100 with symmetric label noise from different levels. Results with * are token from [15]. The mean and standard deviation are computed over three runs.

Dataset	CIFAR-10			CIFAR-100		
Methods / Noise	Sym-20%	Sym-50%	Sym-80%	Sym-20%	Sym-50%	Sym-80%
CE	86.5±0.6	80.6±0.2	63.7±0.8	57.9±0.4	47.3±0.2	22.3±1.2
MixUp	93.2±0.3	88.2±0.3	73.3±0.3	69.5±0.2	57.1±0.6	34.1±0.6
M-correction*	94.0	92.0	86.8	73.9	66.1	48.2
DivideMix*	95.2	94.2	93.0	75.2	72.8	58.3
DivideMix	95.6±0.1	94.6±0.1	92.9±0.3	75.3±0.1	72.7±0.6	56.4±0.3
ELR+	94.9±0.2	93.6±0.1	90.4±0.2	75.5±0.2	71.0±0.2	50.4±0.8
Ours (Semi)	**95.9±0.1**	**95.1±0.2**	**93.1±0.2**	**77.4±0.3**	**74.3±0.6**	**61.6±0.6**

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

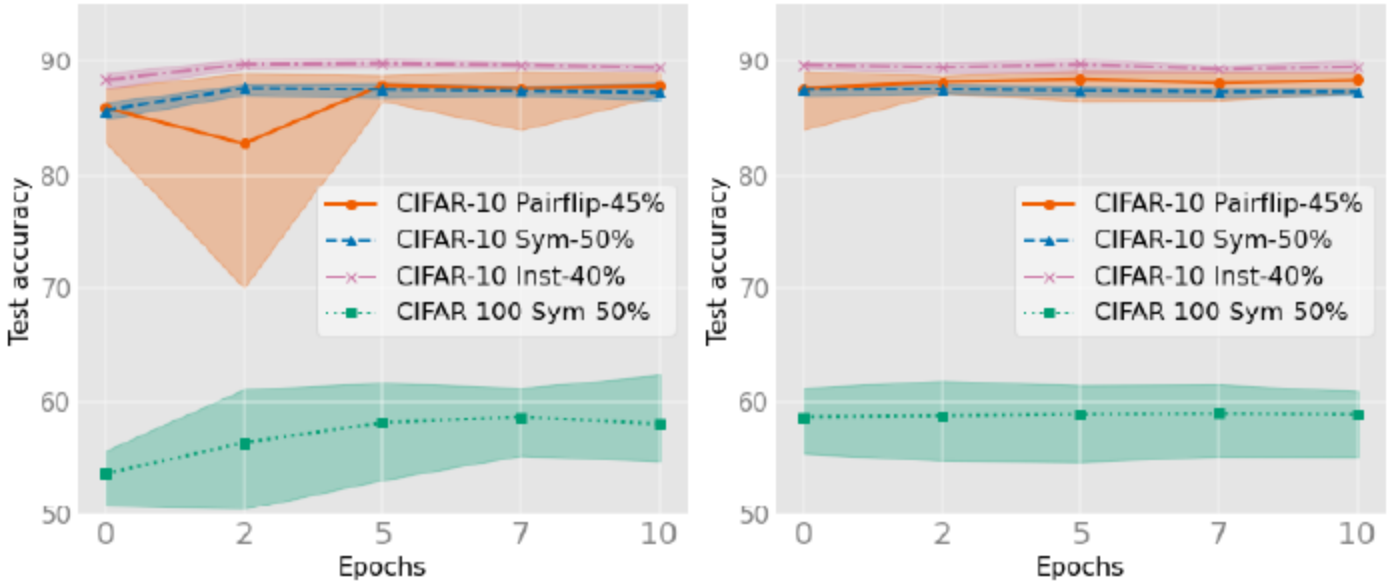
Table 4: Comparison with state-of-the-art methods with semi-supervised learning on CIFAR-10 and CIFAR-100 with instance-dependent and pairflip label noise from different levels. The mean and standard deviation are computed over three runs.

Dataset	CIFAR-10			CIFAR-100		
Methods / Noise	Inst-20%	Inst-40%	Pair-45%	Inst-20%	Inst-40%	Pair-45%
CE	87.5±0.5	78.9±0.7	74.9±1.7	56.8±0.4	48.2±0.5	38.5±0.6
MixUp	93.3±0.2	87.6±0.5	82.4±1.0	67.1±0.1	55.0±0.1	44.2±0.5
DivideMix	95.5±0.1	94.5±0.2	85.6±1.7	75.2±0.2	70.9±0.1	48.2±1.0
ELR+	94.9±0.1	94.3±0.2	86.1±1.2	75.8±0.1	74.3±0.3	65.3±1.3
Ours (Semi)	**95.9±0.1**	**95.3±0.1**	**94.5±0.3**	**77.6±0.3**	**76.1±0.4**	**73.6±1.7**

Table 5: Compassion with state-of-the-art methods on Clothing-1M. Results of baseline methods are taken from the original papers. ours represent the results obtained by PES with a single network and ours* indicate the results obtained by PES with an ensemble model.

CE	Forward	Joint-Optim	DMI	T-revision	DivideMix*	ELR+*	Ours	Ours*
69.21	69.84	72.16	72.46	74.18	74.76	74.81	**74.64**	**74.99**

Experiments(Sensitivity Analysis)



(a) T_2 (b) T_3

Figure 3: Sensitivity analysis for different training iteration numbers: T_2 and T_3.

Do We Need Zero Training Loss After Achieving Zero Training Error?

ICML 2020

Overview

Overfitting

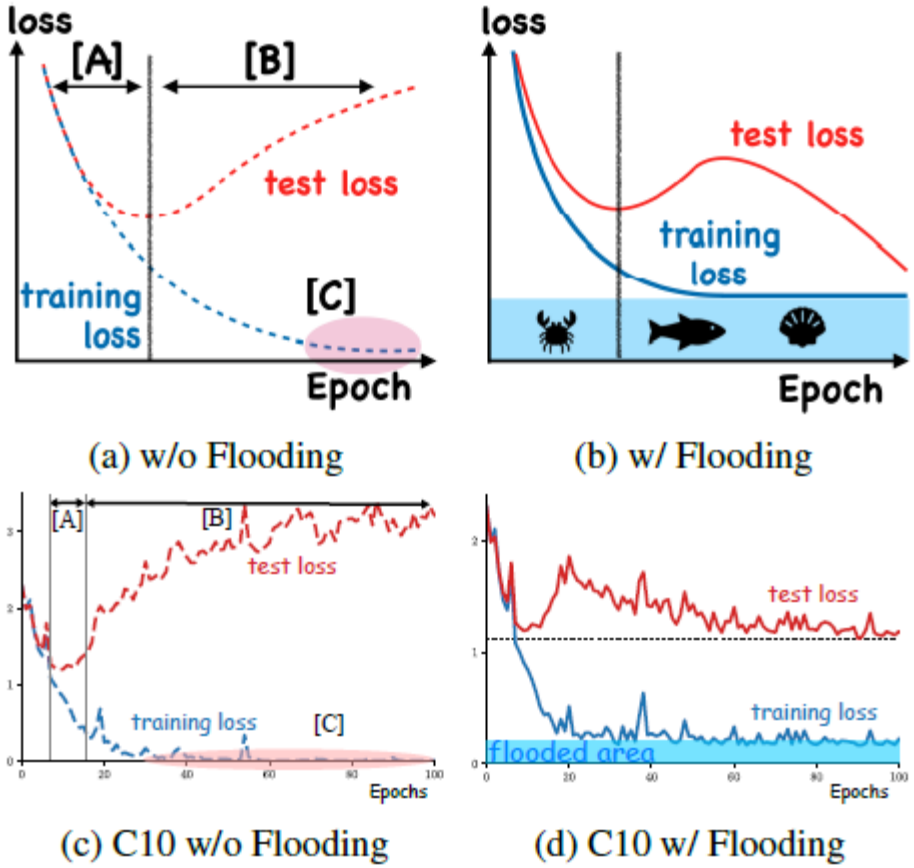
- generalization gap

Zero training error

Zero training loss

Objective function

$$\widetilde{J}(\theta) = |J(\theta) - b| + b,$$



(a) w/o Flooding (b) w/ Flooding

(c) C10 w/o Flooding (d) C10 w/ Flooding

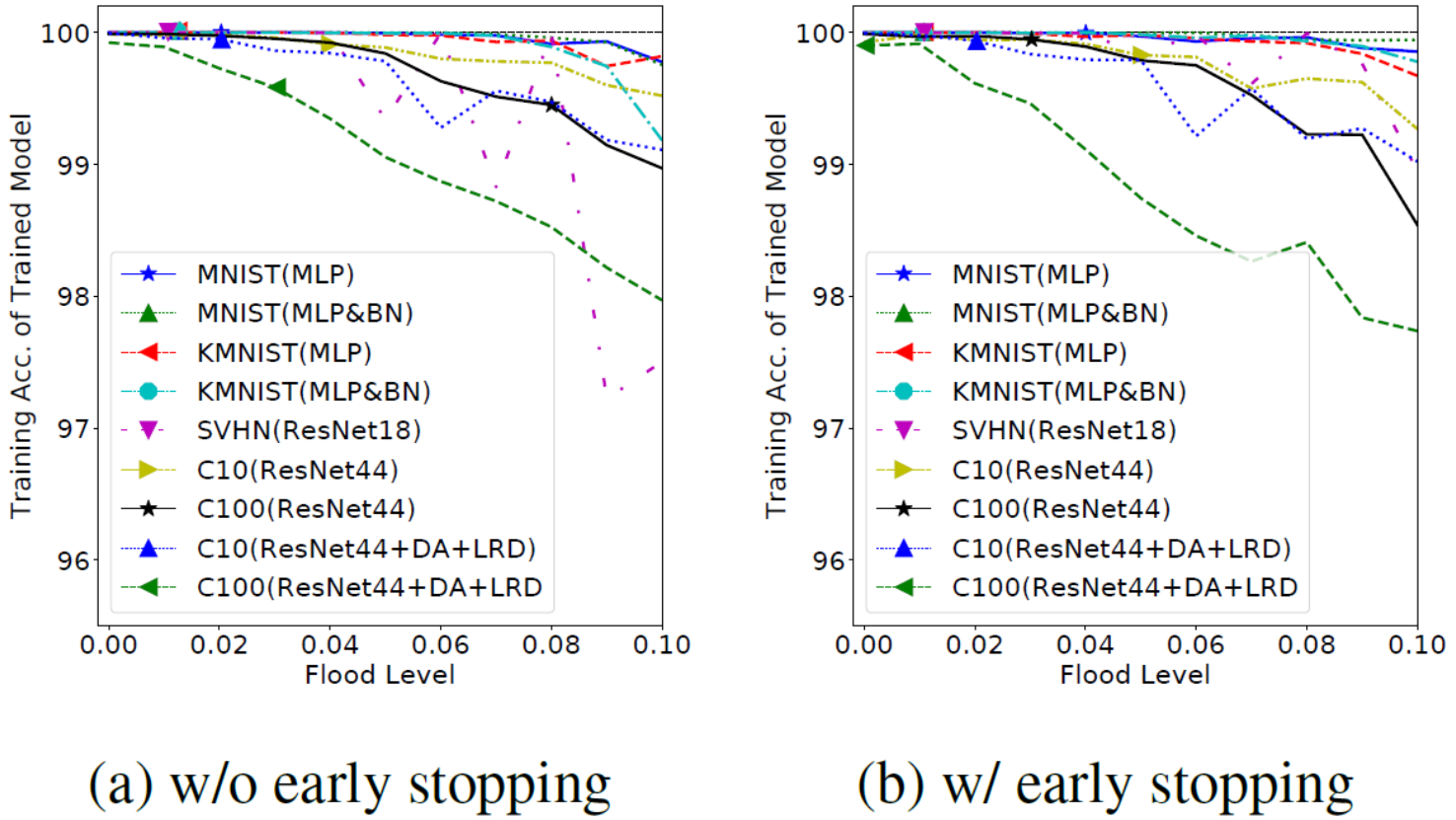
Experiments

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Table 3. Benchmark datasets. Reporting accuracy for all combinations of early stopping and flooding. We compare "w/o flood" and "w/ flood" and the better one is shown in **boldface**. The best setup for each dataset is shown with **underline**. "–" means that flood level of zero was optimal. "LR" stands for learning rate and "aug." is an abbreviation of augmentation.

		w/o early stopping		w/ early stopping	
Dataset	Model & Setup	w/o flood	w/ flood	w/o flood	w/ flood
MNIST	MLP	98.45%	**98.76%**	98.48%	**98.66%**
	MLP w/ weight decay	98.53%	**98.58%**	98.51%	**98.64%**
	MLP w/ batch normalization	98.60%	**98.72%**	**98.66%**	98.65%
Kuzushiji	MLP	92.27%	**93.15%**	92.24%	**92.90%**
	MLP w/ weight decay	92.21%	**92.53%**	92.24%	**93.15%**
	MLP w/ batch normalization	92.98%	**93.80%**	92.81%	**93.74%**
SVHN	ResNet18	92.38%	**92.78%**	92.41%	**92.79%**
	ResNet18 w/ weight decay	93.20%	–	92.99%	**93.42%**
CIFAR-10	ResNet44	**75.38%**	75.31%	74.98%	**75.52%**
	ResNet44 w/ data aug. & LR decay	88.05%	**89.61%**	88.06%	**89.48%**
CIFAR-100	ResNet44	**46.00%**	45.83%	**46.87%**	46.73%
	ResNet44 w/ data aug. & LR decay	63.38%	**63.70%**	63.24%	–

Experiments



(a) w/o early stopping (b) w/ early stopping

Figure 3. Vertical axis is the training accuracy and the horizontal axis is the flood level. Marks are placed on the flood level that was chosen based on validation accuracy.

Thanks