Late Stopping Supplementary

Suqin Yuan[1] Lei Feng[2] Tongliang Liu[1]
[1]The University of Sydney [2]Nanyang Technological University

A1. Datasets and generating label noise

Datasets As shown in Table 1, we verify the effectiveness of our proposed approach on three datasets, i.e., CIFAR-10, CIFAR-100 [4], and CIFAR-10N [6].

Table 1. Summarized information of datasets in our experiments.

Datasets	Train / Test Size	Classes	Noise level
CIFAR-10	$50k$ / $10k$	10	20% / 40%
CIFAR-100	$50k$ / $10k$	100	20% / 40%
CIFAR-10N (*Worst*)	$50k$ / $10k$	10	40.21%

Generating label noise Since CIFAR-10 and CIFAR-100 datasets are clean, following [2, 1, 7], we need to corrupt these datasets manually by the noise transition matrix Q, where $Q_{ij} = Pr(\tilde{y} = j|y = i)$ given that noisy $\tilde{y}$ is flipped from clean y. We assume that the matrix Q has two representative structures:

(1) Symmetric label noise [5]: As shown in Figure 1, if the flip rate is a, the diagonal entries of a symmetric transition matrix are $1 - a$ and the off-diagonal entries are $a/(c - 1)$, where c denotes the number of categories.

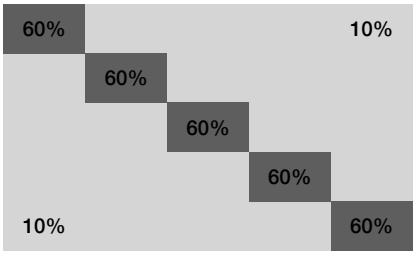

Figure 1. Transition matrices of Symmetric class-dependent label noise, i.e., *Sym. 40%* (using 5 classes as an example).

(2) Instance-dependent label noise: As demonstrated in Algorithm 1, we employ the same construction algorithm as [1] to estimate the instance-dependent label noise by exploiting part-dependent label noise [8]. The actual flip rate is contingent upon both the pre-setting noise ratio τ and the representation of images.

Algorithm 1 Instance-dependent Label Noise Generation

Input: Clean samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$; Noise rate τ.
1: Sample instance flip rates $q \in \mathbb{R}^n$ from the truncated normal distribution $\mathcal{N}(\tau, 0.1^2, [0, 1])$;
2: Independently sample $w_1, w_2, \ldots, w_c$ from the standard normal distribution $\mathcal{N}(0, 1^2)$;
3: For $i = 1, 2, \ldots, n$ do
4: $p = \boldsymbol{x}_i \times w_{y_i}$; //generate instance-dependent flip rates
5: $p_{y_i} = -\infty$; //control the diagonal entry of the instance-dependent transition matrix
6: $p = q_i \times softmax(p)$; //make the sum of the off-diagonal entries of the y_i-th row to be q_i
7: $p_{y_i} = 1 - q_i$; //set the diagonal entry to be 1-q_i
8: Randomly choose a label from the label space according to the possibilities p as noisy label $\bar{y}_i$;
9: End for.
Output: Noisy samples $\{(\boldsymbol{x}_i, \bar{y}_i)\}_{i=1}^n$

A2. Network structure and experimental setup

All of our experiments were conducted using PyTorch v1.11.0. To ensure stable predictions of DNNs, we employed a typical warming-up strategy. Specifically, for CIFAR-10 (N) and CIFAR-100, we trained the network for 4 and 6 epochs, respectively, during the warming-up stage. In our experiments, we applied typical data augmentations such as horizontal flipping and random cropping. For the optimizer settings, we used stochastic gradient descent (SGD) with a momentum of 0.9, weight decay of 5e-4, and a batch size of 128.

During the final round of training, we set the initial learning rate to 0.02 and divided it by 2 in epoch $\{10, 20, 40, 60, 80\}$. All networks were trained for 300 epochs in the final round. For earlier rounds of training, we initialized the learning rate to 0.01 and terminated each round based on the number of FkL-examples for that particular round. For the experiments on synthetic noisy datasets, ResNet-18 and ResNet-34 networks [3] are used for CIFAR-10 and CIFAR-100, respectively. For the experiments on real-world noisy datasets, ResNet-34 networks are used for CIFAR-10N.

To better observe the intrinsic robust learning ability of

Table 2. Summarized information of round number of Late Stopping and k of FkL for different datasets in our experiments.

Training set	$k=1$	$k=2$	$k=3$
CIFAR-10 (*20% noise*)	4	4	4
CIFAR-10 (*40% noise*)	7	7	7
CIFAR-10N (*Worst*)	7	7	7
CIFAR-100 (*20% noise*)	0	8	4
CIFAR-100 (*40% noise*)	22	5	0

DNNs, we varied the round number of Late Stopping and value of k in FkL for different datasets in our experiments. As illustrated in Table 2, we adjusted the round number of Late Stopping based on the noise level of the datasets. The value of k in FkL is defined as the number of epochs required for an example to be first-time consistently and correctly classified. In a similar fashion, we varied the value of k during the training process by incrementally increasing it from a low value to a high value.

B1. Experiments on falsely retained examples

In this section, we conducted experiments with *falsely retained examples* collected from the final training set of a CIFAR-10 (*Sym.40%*) dataset where Late Stopping was applied. We refer to these examples as S_w. To evaluate the hardness of *falsely retained examples* before and after being mislabeled, we collected the rankings of S_w in the *Sym.40%* noise dataset based on *loss* criterion and FkL criterion after the first round of Late Stopping. Meanwhile, we collected the rankings of S_w in the clean dataset based on the same criteria after the first round of Late Stopping.

Table 3. The comparison of the average ranking of *false retaing examples* using the *loss* criterion and the FkL criterion before and after fixing the noisy labels (CIFAR-10, Before: *Sym. 40%* noise).

Label	Criterion	*Avg.* Ranking
Before fixing (Given label)	FkL	22340.66
	loss	23673.50
After fixing (Ground-truth label)	FkL	28452.08 (**+27.36%**)
	loss	29476.46 (**+24.51%**)

Table 3 presents the results of our experiments, which indicate that *falsely retained examples* are significantly easier for the classifier to learn after being mislabeled based on the *loss* criterion and FkL criterion. Specifically, comparing the rankings of S_w in the noisy dataset and the clean dataset, we observed that the hardness of these examples decreases after being mislabeled.

References

[1] Yingbin Bai and Tongliang Liu. Me-momentum: Extracting hard confident examples from noisily labeled data. In *ICCV*, 2021.

[2] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *NeurIPS*, 2018.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[4] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[5] Brendan Van Rooyen, Aditya Menon, and Robert C Williamson. Learning with symmetric label noise: The importance of being unhinged. *NeurIPS*, 2015.

[6] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. *arXiv*, 2021.

[7] Qi Wei, Haoliang Sun, Xiankai Lu, and Yilong Yin. Self-filtering: A noise-aware sample selection for label noise with confidence penalization. In *ECCV*, 2022.

[8] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. *NeurIPS*, 2020.