Appendix for ”Self-Filtering: A Noise-Aware Sample Selection for Label Noise with Confidence Penalization”

Qi Wei, Haoliang Sun, Xiankai Lu, and Yilong Yin

School of Software, Shandong University, Jinan, China
`{1998v7,haolsun.cn,carrierlxk}@gmail.com, ylyin@sdu.edu.cn`

1 The generation of noisy labels

Reference to [3], we manually generate the instance-dependent label noise according to the following algorithm. The illustration of other types of label noise is shown in Figure 1.

Algorithm 1 Instance-dependent Label Noise Generation

Input: Clean set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$; 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: $p = x_i \times w_{y_i}$; //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 $\{(x_i, \bar{y}_i)\}_{i=1}^{n}$

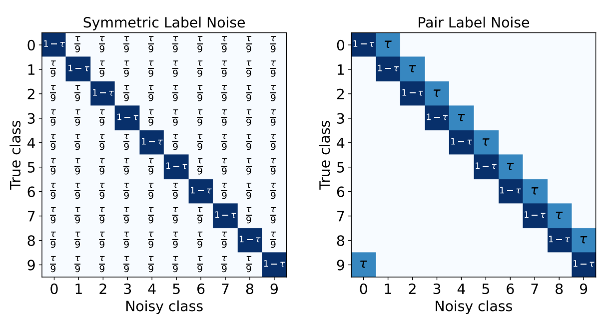


Fig. 1: Symmetric and pair-flipped label noise on CIFAR-10 as noise ratio is τ.

2 Combination SFT with FixMatch

Specially, we filter the examples with fluctuation events in the memory bank and remove their labels to construct an unlabeled set $\mathcal{D}_u = \{\mathbf{u}^i \in \mathcal{D} | \beta^i = 1\}_{i=1}^{n}$. Following the operation in FixMatch, we first compute the prediction vector $\mathbf{q} = f(\mathbf{u}', \theta)$ of the weakly-augmented version of sample $\mathbf{u}'$ and set $\tilde{y} =_j (\mathbf{q}_j)$ as the pseudo label. Then, we conduct a strong augmentation $\mathcal{A}^s(\mathbf{u})$ for $\mathbf{u}$. The objective function of semi-supervised learning (SSL) can be written as

$$\mathcal{L}_{ssl} = \mathbb{1}(\mathbf{q}_{\tilde{y}} \geq c)\mathcal{L}_{\mathrm{CE}}(\tilde{y}, f(\mathcal{A}^s(\mathbf{u}), \theta)), \tag{1}$$

where c is a confidence coefficient.

By combining the SSL loss with the training loss in SFT, we can obtain the final objective function

$$\begin{aligned} \mathcal{L} = \mathbb{E}_{(\mathbf{x},y)\in\tilde{\mathcal{D}}}[\mathcal{L}_{\mathrm{CE}}(f(\mathbf{x}, \theta), y) + \lambda\mathcal{L}_{\mathrm{CR}}(f(\mathbf{x}, \theta), y)] \\ + \mathbb{E}_{(\mathbf{u},y)\in\mathcal{D}_u}[\mathcal{L}_{ssl}(f(\mathbf{u}, \theta), \tilde{y})]. \end{aligned} \tag{2}$$

3 More analysis of confidence penalty $\mathcal{R}$

Comparisons with Entropy regularization. Entropy regularization (ER) [2] is regarded as a means of improving classifier confidence with restraining minimum entropy. [1] tackles the issue that warm-up is not effective for pair label noise by adding the negative entropy to the negative log-likelihood: $\mathcal{L}(\theta) = -\sum logp_\theta(y|x) - \beta H(p_\theta(y|x))$. To the same target, we propose a confidence regularization term $\mathcal{R}$ to mitigate overconfidence.

To verify that our method outperforms ER in confronting label noise, we conduct the experiments under variant noise conditions by replacing $\mathcal{R}$ with ER. The result is shown in Table 1. $\mathcal{R}$ attains the greater superiority compared with ER as the noise ratio increases.

Table 1: Improvement with regularization on CIFAR-10.

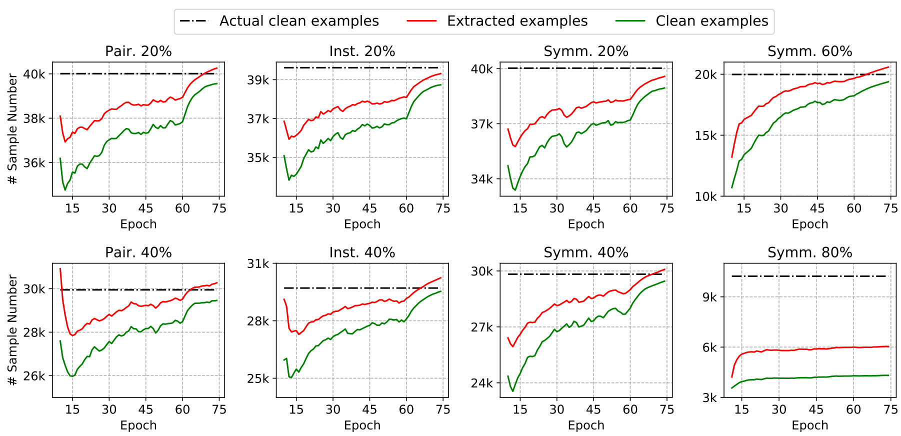
Methods	**I** 20%	**I** 40%	**I** 60%	**P** 10%	**P** 20%	**P** 30%	**P** 40%
ER	91.24	89.46	86.22	92.04	91.53	90.78	89.93
Ours	+0.17	+0.51	+1.92	+0.20	+0.29	+0.54	+0.76

Effects on selecting. We would clarify that the regulizer $\mathcal{R}$ can help maintain boundary examples to some degree. Hence, we discard the regulizer and compute F1-scores of selection results in Table 2. Our *fluctuation* criterion consistently achieves fairly higher F1-scores compared to small-loss criterion. The effect of regularizers is marginal for selecting clean samples.

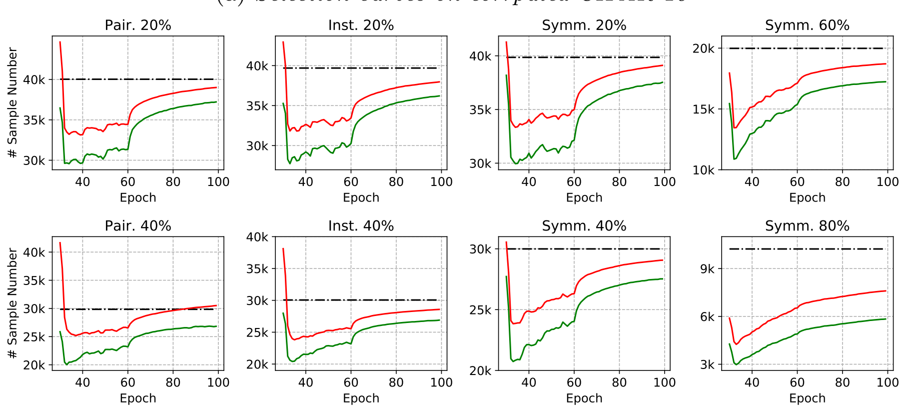
4 Additional experiment results

Table 2: F1-scores of selection results on CIFAR-10.

Methods	**S** 40%	**S** 80%	**P** 40%	**I** 40%
Small loss	0.881±0.008	0.326±0.014	0.834±0.010	0.810±0.015
Ours w/o Regu.	0.969±0.006	0.581±0.008	0.930±0.006	0.957±0.010
Ours w Regu.	0.986±0.002	0.607±0.009	0.954±0.006	0.967±0.004

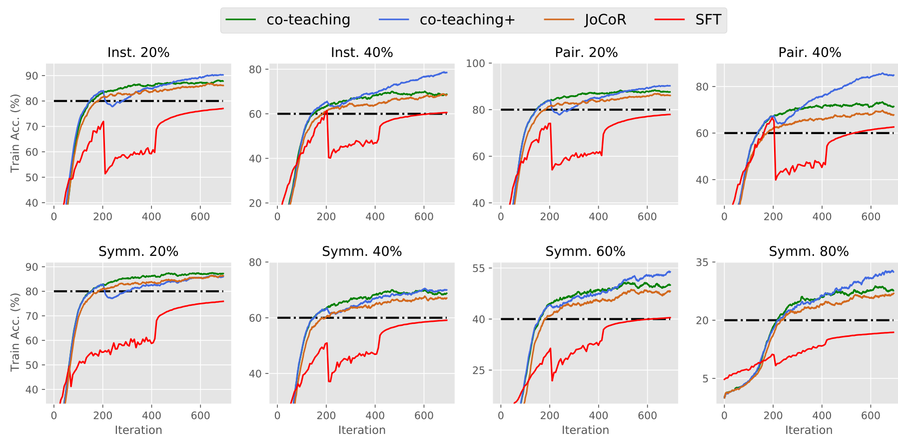


(a) *Selection curves on corrputed CIFAR-10*

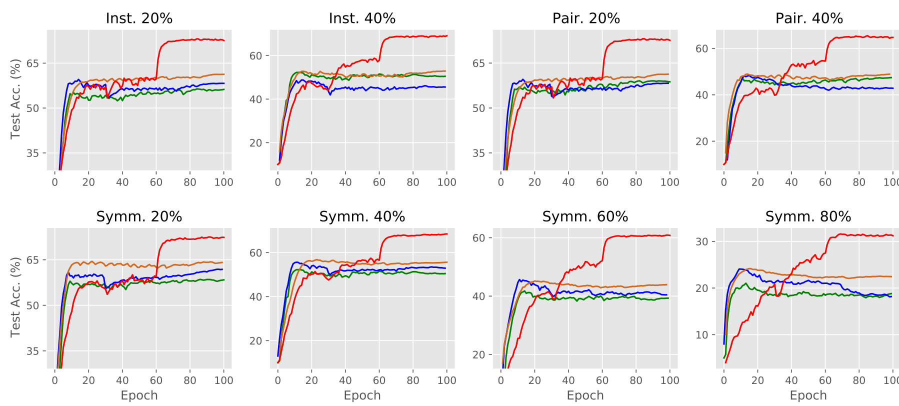


(b) *Selection curves on corrputed CIFAR-100*

Fig. 2: The clean labels number in extracted confident examples. The black dashed line denotes the number of the actual clean labels. SFT achieves more than 95% selection precision in majority of noise conditions.



(a) *Train accuracy (%) on CIFAR-100*



(b) *Test accuracy (%) on CIFAR-100*

Fig. 3: Comparisons with Train accuracy (%) and test accuracy (%) on CIFAR-100. 1 iteration is set as 50 batches. Especially, as the training proceeds, the network trained with the other three methods exhibits overconfidence in the corrupted training set. Namely, their training curve exceeds the black horizontal line while SFT is always below this line. The results indicate the robust learning of SFT on the corrupted dataset and demonstrate the effectiveness of our algorithm in learning with noisy labels.

References

1. Li, J., Socher, R., Hoi, S.C.: Dividemix: Learning with noisy labels as semi-supervised learning. In: ICLR (2020) 2
2. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., Hinton, G.: Regularizing neural networks by penalizing confident output distributions. In: ICLR (2017) 2
3. Xia, X., Liu, T., Han, B., Wang, N., Gong, M., Liu, H., Niu, G., Tao, D., Sugiyama, M.: Part-dependent label noise: Towards instance-dependent label noise. In: NeurIPS (2020) 1