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

Qi Wei, Haoliang Sun, Xiankai Lu, Yilong Yin *TIME Lab, Shandong University*

Towards Intelligence MEchanism
ECCV TEL AVIV 2022

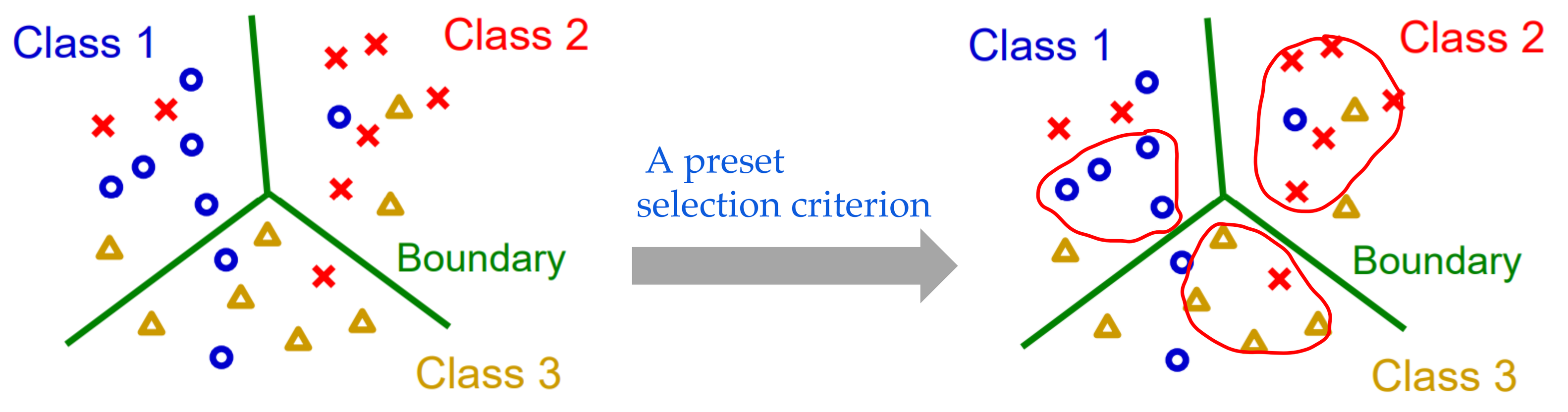
Summary

- A novel selection criterion dubbed fluctuation criterion is proposed for retaining valuable samples lying around decision boundary.
- A confidence regularization term is designed to further mitigate the over-confidence in noisy samples.
- Any semi-supervised method can be applicable to our framework, improving the performance of SFT.
- SFT outperforms its counterparts by sharp margins.

Code is available at *https://github.com/1998v7/Self-Filtering*

Sample selection strategy

Main idea: Use a preset selection criterion to select a subset with smaller noise ratio from the label-corrupted training set.

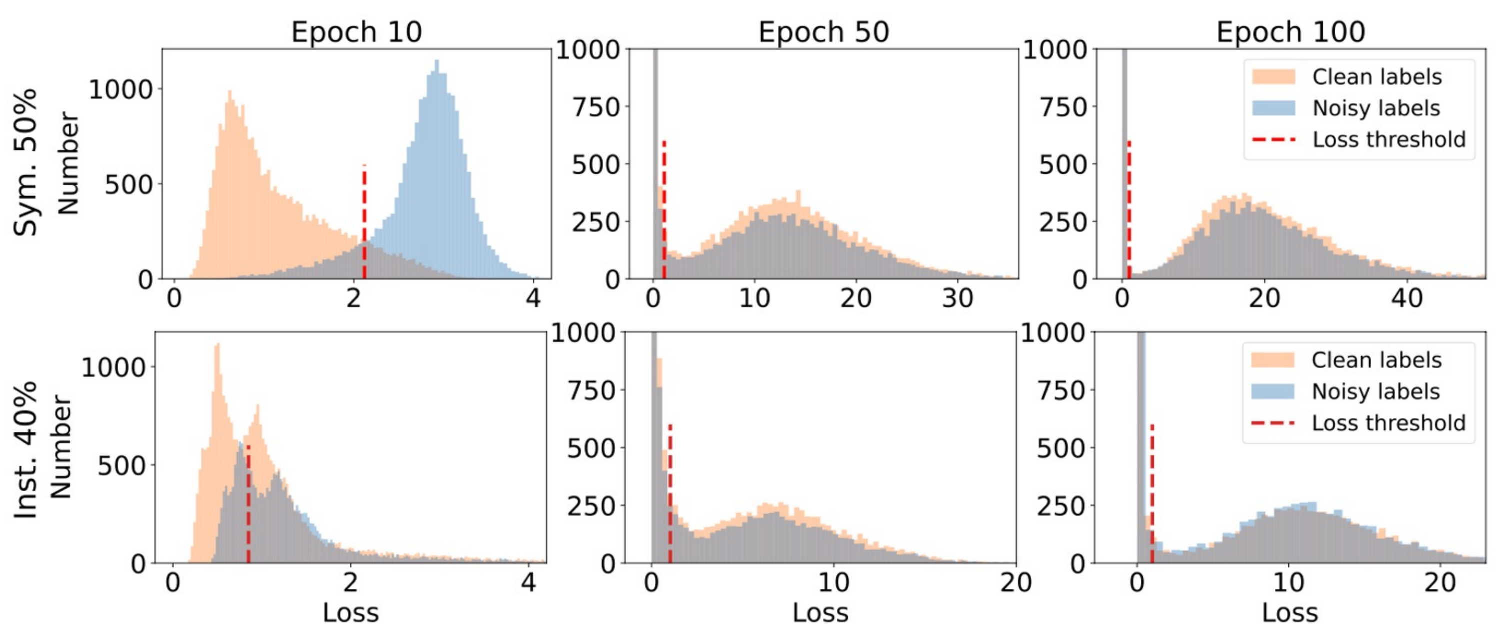


A label-corrupted training set   A subset with smaller noise ratio

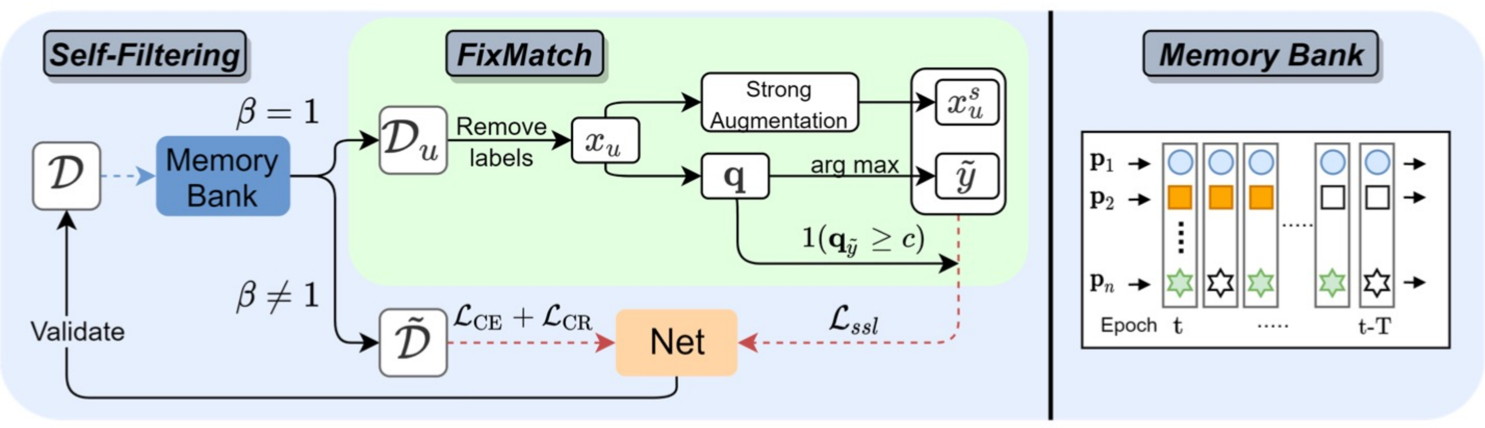
ACML 2021 Tutorial. Learning under Noisy Supervision

Selection bias in the small-loss criterion

Essential boundary samples are entangled with noise samples and discarded.



Our framework *Self-Filtering*



(a) Flowchart of Self-Filtering+   (b) Illustration of memory bank

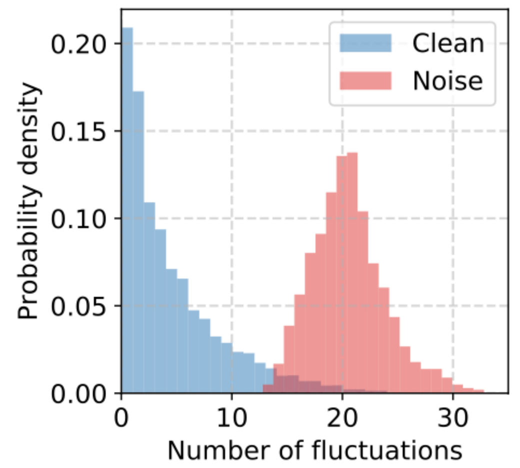
Fluctuation selection criterion

- Definition of **fluctuation event**:

$$\beta = (\arg\max(p^{t_1}) = y) \wedge (\arg\max(p^{t_2}) \neq y)$$

- The selected set (filter the noise)

$$\widetilde{D} = \{(x^i, y^i) \in D^{train} | \beta^i \neq 1\}_{i=1}^{N}$$



The fluctuation criterion provides discriminative information for filtering the noise as shown in right figure.

Confidence regularization

- An adaptive weight function

$$\alpha(p_j) = \max(0, \mathrm{T} - \frac{p_j}{p_y})$$

- Confidence regularization term

$$L_{CR} = -\frac{1}{K}\sum_{k\in[K]} \alpha(p_j) \cdot \log p_k$$

Merits:

- muting at the beginning and casting the objective to cross entropy for **fast convergence.**
- **Adaptive** strength for confidence penalty

Improved by semi-supervised technique

Self-Filtering can be improved by current semi-supervised learning strategy.

The selected (clean) set: $\widetilde{D} = \{(x^i, y^i) \in D^{train} | \beta^i \neq 1\}_{i=1}^{N}$

The filtered (noisy) set: $\widehat{D} = \{(x^i, y^i) \in D^{train} | \beta^i = 1\}_{i=1}^{N}$

For $(x, y) \in \widetilde{D}$ and $(x', y') \in \widehat{D}$, the total training objective:

$$L_{CR}(x, y) + \alpha \cdot L_{SSL}(x', y')$$

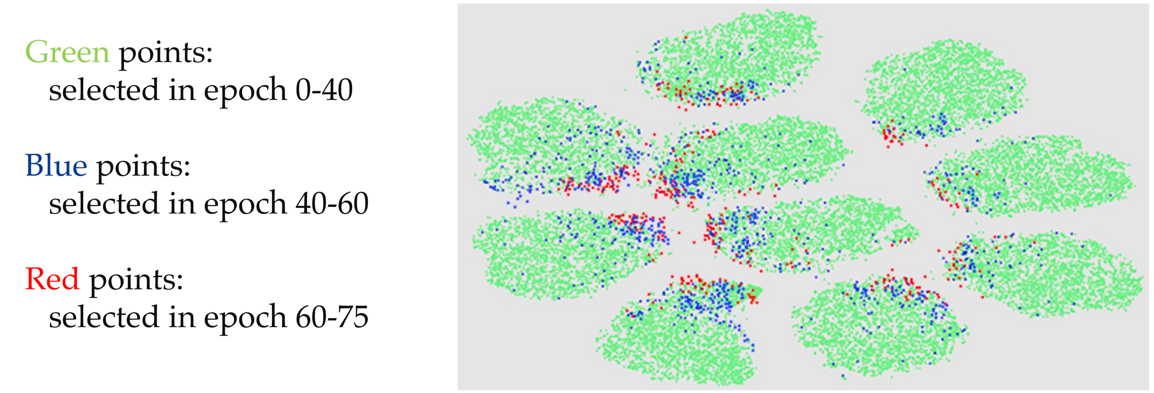
Experimental results

SFT achieves the SOTA performance on CIFAR-10 and CIFAR-100.

Method	Symm. 20%	Symm. 40%	Pair. 20%	Pair. 40%	Inst. 20%	Inst. 40%
DMI [39]	88.18±0.36	83.98±0.48	89.44±0.41	84.37±0.78	89.14±0.36	84.78±1.97
Peer Loss [17]	88.97±0.47	84.29±0.52	89.61±0.66	85.18±0.87	89.94±0.51	85.77±1.19
Co-teaching [9]	87.16±0.11	83.59±0.28	86.91±0.37	82.77±0.57	86.54±0.11	80.98±0.39
JoCoR [32]	88.69±0.19	85.44±0.29	87.75±0.46	83.91±0.49	87.31±0.27	82.49±0.57
SELFIE [27]	90.18±0.25	86.27±0.31	89.29±0.19	85.71±0.30	89.24±0.27	84.16±0.44
CDR [35]	89.68±0.38	86.13±0.44	89.19±0.29	85.79±0.41	90.24±0.39	83.07±1.33
Me-Momentum [3]	91.44±0.33	88.39±0.34	90.91±0.45	87.49±0.56	90.86±0.21	86.66±0.91
PES[4]	92.38±0.41	87.45±0.34	91.22±0.42	89.52±0.91	**92.69±0.42**	89.73±0.51
SFT (ours)	**92.57±0.32**	**89.54±0.27**	**91.53±0.26**	**89.93±0.47**	91.41±0.32	**89.97±0.49**
DMI [39]	58.73±0.70	49.81±1.22	59.41±0.69	48.13±0.52	58.05±0.20	47.36±0.68
Peer Loss [17]	58.41±0.55	50.53±1.31	58.73±0.51	50.17±0.42	58.91±0.41	48.61±0.78
Co-teaching [9]	59.28±0.47	51.60±0.49	58.07±0.61	49.79±0.69	57.24±0.69	49.39±0.99
JoCoR [32]	64.17±0.19	55.97±0.46	60.42±0.35	50.97±0.58	61.98±0.39	50.59±0.71
SELFIE [27]	67.19±0.30	61.29±0.39	65.18±0.23	58.67±0.51	65.44±0.43	53.91±0.66
CDR [35]	66.52±0.24	60.18±0.22	66.12±0.31	59.49±0.47	67.06±0.50	56.86±0.62
Me-Momentum [3]	68.03±0.53	63.48±0.72	68.42±0.19	59.73±0.47	68.11±0.57	58.38±1.28
PES [4]	68.89±0.41	64.90±0.57	69.31±0.25	59.08±0.81	70.49±0.72	65.68±0.44
SFT (ours)	**71.98±0.26**	**69.72±0.31**	**71.23±0.29**	**69.29±0.42**	**71.83±0.42**	**69.91±0.54**

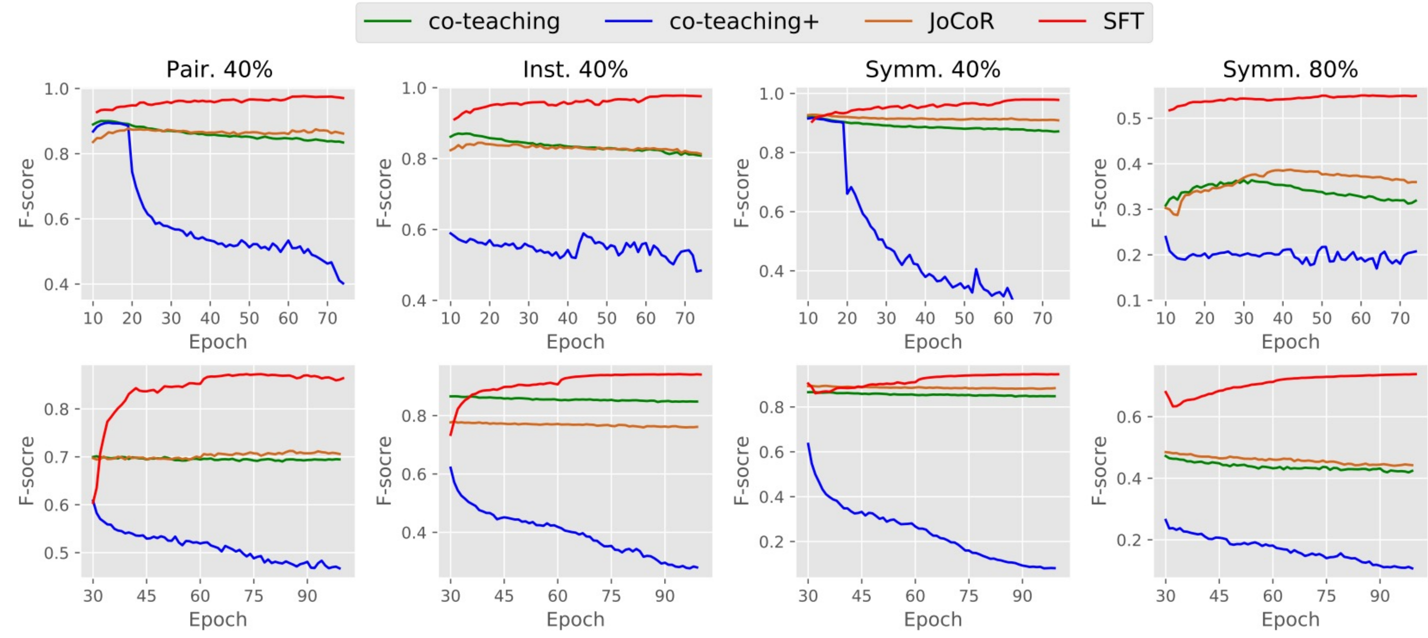
More analyses

- Visualization of selection



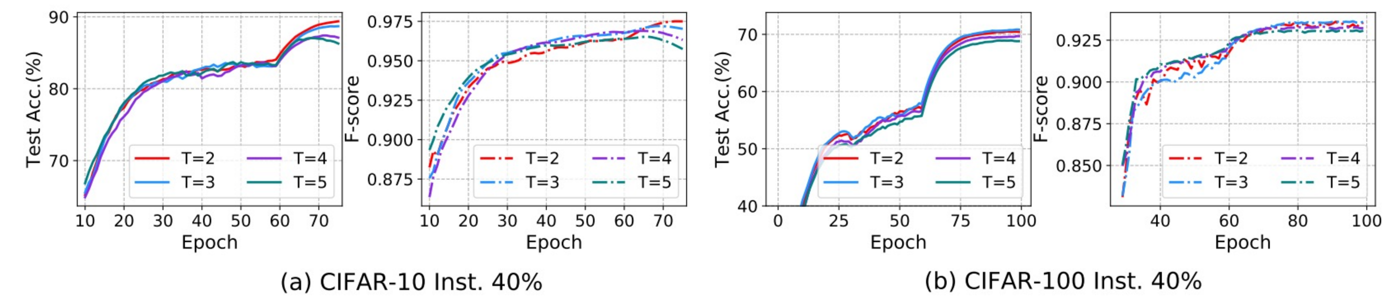
SFT selects more boundary examples as training proceeds.

- Stable selection curves



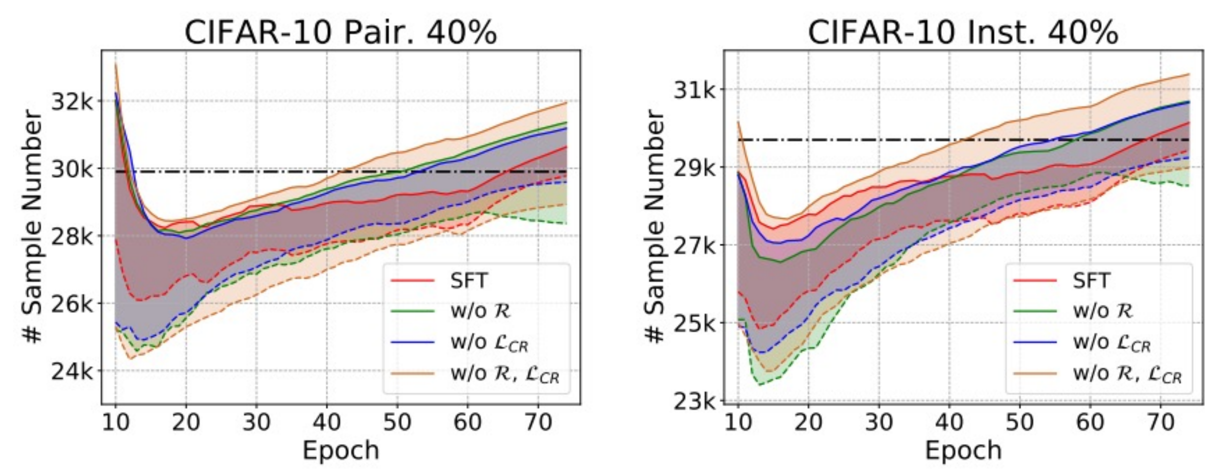
Higher F1-score of selection results is achieved by *SFT*.

- Hyper-parameter selection



SFT is not sensitive to hyper-parameters.

- Ablation study



With the support of the two terms, the selected subset contains less noisy labels

Contact

Mail: 1998v7@gmail.com
Towards Intelligence Mechanism Lab
School of Software – Shandong University – China