南京航空航天大學
Nanjing University of Aeronautics and Astronautics

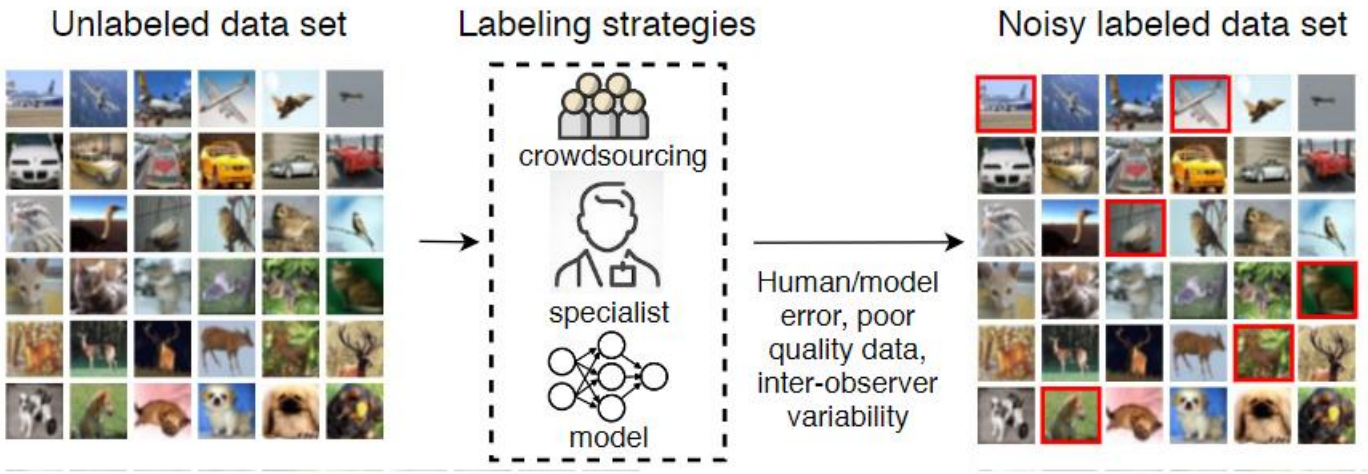
Class2Simi: A Noise Reduction Perspective on Learning with Noisy Labels

Songhua Wu [*1] **Xiaobo Xia** [*1] **Tongliang Liu** [1]
Bo Han [2] **Mingming Gong** [3] **Nannan Wang** [4] **Haifeng Liu** [5] **Gang Niu** [6]

ICML 2021

Motivation

Labeling process and noise sources



Almost all existing methods deal with the label noise problem in pointwise manners.

Methods employing pairwise manners are very prevailing and have made a great success in machine learning(Contrastive Learning)

Does learning in a pairwise manner mitigate label noise?

Overview

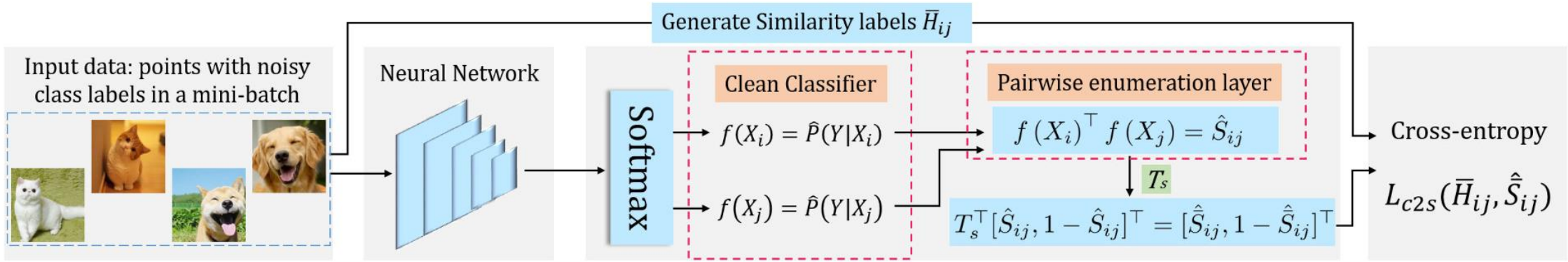
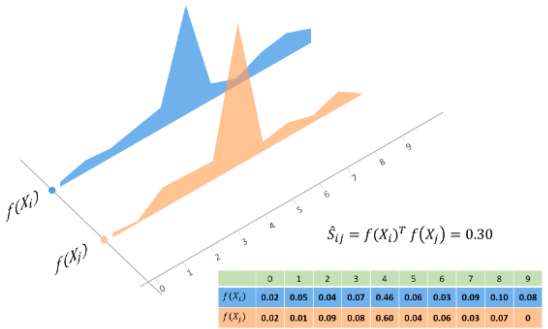


Figure 2. An overview of the proposed method. We add a pairwise enumeration layer and similarity transition matrix to calculate and correct the predicted similarity posterior. By minimizing the proposed loss L_{c2s}, a classifier f can be learned for assigning clean labels.

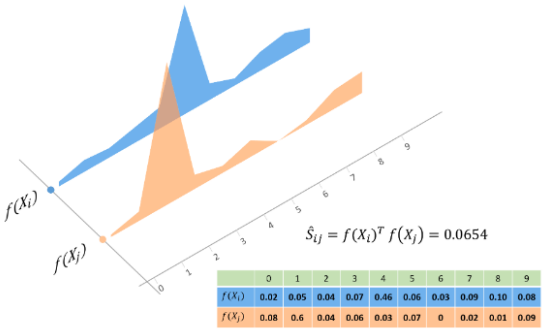
Core issues:

1. Transformation on labels and the transition matrix
2. Learning with noisy similarity labels



	0	1	2	3	4	5	6	7	8	9
$f(X_i)$	0.02	0.05	0.04	0.07	0.46	0.06	0.03	0.09	0.10	0.08
$f(X_j)$	0.02	0.01	0.09	0.08	0.60	0.04	0.06	0.03	0.07	0

(a) *Similar example*



	0	1	2	3	4	5	6	7	8	9
$f(X_i)$	0.02	0.05	0.04	0.07	0.46	0.06	0.03	0.09	0.10	0.08
$f(X_j)$	0.08	0.6	0.04	0.06	0.03	0.07	0	0.02	0.01	0.09

(b) *Dissimilar example*

Transformation on labels

x:	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
y:	1	1	2	2	3	3	4	4

Instances with clean class labels

x:	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$\bar{y}$:	1	2	2	3	3	4	4	1

Instances with noisy class labels

	x_1 1	x_2 1	x_3 2	x_4 2	x_5 3	x_6 3	x_7 4	x_8 4
x_1	1	1	0	0	0	0	0	0
x_2	1	1	0	0	0	0	0	0
x_3	0	0	1	1	0	0	0	0
x_4	0	0	1	1	0	0	0	0
x_5	0	0	0	0	1	1	0	0
x_6	0	0	0	0	1	1	0	0
x_7	0	0	0	0	0	0	1	1
x_8	0	0	0	0	0	0	1	1

Data pairs with clean similarity labels

	x_1 1	x_2 2	x_3 2	x_4 3	x_5 3	x_6 4	x_7 4	x_8 1
x_1	1	0	0	0	0	0	0	1
x_2	0	1	1	0	0	0	0	0
x_3	0	1	1	0	0	0	0	0
x_4	0	0	0	1	1	0	0	0
x_5	0	0	0	1	1	0	0	0
x_6	0	0	0	0	0	1	1	0
x_7	0	0	0	0	0	1	1	0
x_8	1	0	0	0	0	0	0	1

Data pairs with noisy similarity labels

	1	2	3	4
1	0.5	0.5		
2		0.5	0.5	
3			0.5	0.5
4	0.5			0.5

Class transition matrix

	0	1
0	5/6	1/6
1	1/2	1/2

Similarity transition matrix

Similarity labels:

if the two instances have the same class label, assign this pair a similarity label 1, otherwise 0.

Dim(Similarity transition matrix) = 2x2

T_S similarity transition matrix

T_c class transition matrix

Theorem 1 *Assume that the dataset is balanced (each class has the same amount of instances, and c classes in total), and the noise is class-dependent. Given a class transition matrix T_c, such that $T_{c,ij} = P(\bar{Y} = j|Y = i)$. The elements of the corresponding similarity transition matrix T_s can be calculated as*

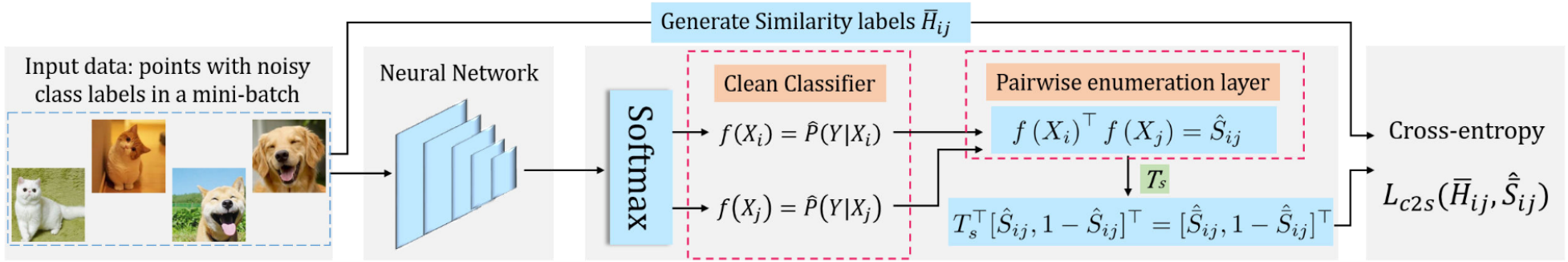
$$T_{s,00} = \frac{c^2 - c - \left(\sum_j (\sum_i T_{c,ij})^2 - ||T_c||_{\text{Fro}}^2\right)}{c^2 - c},$$

$$T_{s,01} = \frac{\sum_j (\sum_i T_{c,ij})^2 - ||T_c||_{\text{Fro}}^2}{c^2 - c},$$

$$T_{s,10} = \frac{c - ||T_c||_{\text{Fro}}^2}{c}, \qquad T_{s,11} = \frac{||T_c||_{\text{Fro}}^2}{c}.$$

Theorem 2 *Assume that the dataset is balanced (each class has the same amount of samples), and the noise is class-dependent. When the number of classes $c \geq 8$, the noise rate of noisy similarity labels is lower than that of the noisy class labels.*

Learning with noisy similarity labels



The predicted clean similarity posterior: $\hat{S}_{ij} = f(X_i)^\top f(X_j)$

Clean similarity posterior: $P(H_{ij} \mid X_i, X_j)$

Noisy similarity posterior: $P(\bar{H}_{ij} \mid X_i, X_j)$

$$P(\bar{H}_{ij} \mid X_i, X_j) = T_s^\top P(H_{ij} \mid X_i, X_j)$$

Optimization function:

$$L_{c2s}(\bar{H}_{ij}, \hat{S}_{ij}) = -\sum_{i,j} \bar{H}_{ij} \log \hat{\bar{S}}_{ij} + (1-\bar{H}_{ij}) \log(1-\hat{S}_{ij})$$

x:	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
y:	1	1	2	2	3	3	4	4

Instances with clean class labels

x:	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$\bar{y}$:	1	2	2	3	3	4	4	1

Instances with noisy class labels

		x_1 1	x_2 1	x_3 2	x_4 2	x_5 3	x_6 3	x_7 4	x_8 4
x_1	1	1	1	0	0	0	0	0	0
x_2	1	1	1	0	0	0	0	0	0
x_3	2	0	0	1	1	0	0	0	0
x_4	2	0	0	1	1	0	0	0	0
x_5	3	0	0	0	0	1	1	0	0
x_6	3	0	0	0	0	1	1	0	0
x_7	4	0	0	0	0	0	0	1	1
x_8	4	0	0	0	0	0	0	1	1

Data pairs with clean similarity labels

		x_1 1	x_2 2	x_3 2	x_4 3	x_5 3	x_6 4	x_7 4	x_8 1
x_1	1	1	0	0	0	0	0	0	1
x_2	2	0	1	1	0	0	0	0	0
x_3	2	0	1	1	0	0	0	0	0
x_4	3	0	0	0	1	1	0	0	0
x_5	3	0	0	0	1	1	0	0	0
x_6	4	0	0	0	0	0	1	1	0
x_7	4	0	0	0	0	0	1	1	0
x_8	1	1	0	0	0	0	0	0	1

Data pairs with noisy similarity labels

	1	2	3	4
1	0.5	0.5		
2		0.5	0.5	
3			0.5	0.5
4	0.5			0.5

Class transition matrix

	0	1
0	5/6	1/6
1	1/2	1/2

Similarity transition matrix

Implementation

Algorithm 1 Class2Simi

Input: training data with noisy class labels; validation data with noisy class labels.

Stage 1: Learn $\hat{T}_s$

1: Learn $g(X) = \hat{P}(\bar{Y}|X)$ by training data with noisy class labels, and save the model for Stage 2;

2: Estimate $\hat{T}_c$ following the optimization method in (Patrini et al., 2017);

3: Transform $\hat{T}_c$ to $\hat{T}_s$.

Stage 2: Learn the classifier $f(X) = \hat{P}(Y|X)$

4: Load the model saved in Stage 1, and train the whole pipeline showed in Figure 2.

Output: classifier f.

The expected and empirical risks can be defined as:

$$R(f) = E_{(X_i,X_j,\bar{Y}_i,\bar{Y}_j,\bar{H}_{ij},T_s)\sim\mathcal{D}_\rho}\left[\ell\left(f(X_i), f(X_j), T_s, \bar{H}_{ij}\right)\right]$$

$$R_n(f) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\ell\left(f(X_i), f(X_j), T_s, \bar{H}_{ij}\right)$$

Generalization error bound:

$$R(\hat{f}) - R_n(\hat{f}) \leq M\sqrt{\frac{\log 1/\delta}{2n}} + \frac{(T_{s,11}-T_{s,01})2Bc\left(\sqrt{2d\log 2}+1\right)\prod_{i=1}^{d}M_i}{T_{s,11}\sqrt{n}}$$

Experiment

Table 1. Means and Standard Deviations of Classification Accuracy over 5 trials on image datasets.

MNIST	Sym-0.2	Sym-0.4	Sym-0.6	Asym-0.2	Asym-0.4	Asym-0.6
Co-teaching	97.34±0.26	94.68±0.52	93.36±0.47	97.37±0.20	96.63±0.41	91.33±0.38
JoCor	97.48±0.12	96.31±0.20	93.18±0.27	97.31±0.09	95.73±0.29	91.43±0.28
PHuber-CE	98.65±0.18	98.17±0.15	97.63±0.36	98.73±0.09	98.36±0.25	97.37±0.41
APL	98.77±0.21	97.06±0.37	97.67±0.35	98.72±0.10	98.45±0.29	97.58±0.25
S2E	98.96±0.27	93.27±2.18	89.37±0.70	99.19±0.05	94.47±1.08	92.36±2.40
Revision	98.92±0.09	98.42±0.50	98.10±0.37	98.97±0.06	98.58±0.19	98.21±0.19
Reweight	98.78±0.16	98.26±0.22	97.02±0.58	98.62±0.19	98.12±0.31	96.98±0.29
Forward	98.76±0.03	98.37±0.25	96.89±0.49	98.61±0.22	98.08±0.33	97.43±0.25
R-Class2Simi	99.04±0.06	98.87±0.06	98.40±0.17	99.06±0.05	98.75±0.08	98.23±0.20
F-Class2Simi	**99.26±0.07**	**99.18±0.06**	**98.91±0.09**	**99.26±0.05**	**99.08±0.07**	**98.91±0.07**
CIFAR10	Sym-0.2	Sym-0.4	Sym-0.6	Asym-0.2	Asym-0.4	Asym-0.6
Co-teaching	88.92±0.45	85.97±1.02	75.97±1.33	89.14±0.36	84.77±1.08	76.07±1.27
JoCor	88.46±0.25	85.19±0.75	77.03±0.92	88.96±0.70	85.19±0.58	75.76±1.31
PHuber-CE	90.37±0.26	86.05±0.37	74.06±0.92	90.73±0.22	86.06±0.53	73.25±1.04
APL	89.07±0.92	85.77±0.84	70.06±1.06	89.97±0.19	85.60±0.91	72.33±1.68
S2E	90.04±1.22	82.05±1.95	57.96±4.70	90.12±0.97	83.16±1.58	64.77±3.06
Revision	90.02±0.48	85.47±0.71	73.92±2.02	89.77±0.28	85.32±1.36	75.24±1.87
Reweight	89.05±0.32	84.60±0.45	74.87±1.18	89.28±0.26	84.61±0.62	72.77±1.91
Forward	89.63±0.20	87.08±0.31	73.24±1.33	90.03±0.41	86.64±0.71	77.41±0.43
R-Class2Simi	90.91±0.26	87.80±0.23	79.19±1.65	91.07±0.21	87.78±0.33	78.56±0.63
F-Class2Simi	**91.38±0.19**	**88.22±0.19**	**79.45±0.53**	**91.24±0.27**	**87.79±0.36**	**79.05±0.56**

Experiment

CIFAR100	Sym-0.2	Sym-0.4	Sym-0.6	Asym-0.2	Asym-0.4	Asym-0.6
Co-teaching	57.14±0.49	52.62±1.03	37.32±1.67	57.82±0.37	51.32±0.83	35.32±1.68
JoCoR	58.32±0.71	51.76±1.07	37.02±1.33	58.61±0.30	49.18±1.05	37.09±1.82
PHuber-CE	57.90±0.31	52.36±0.77	37.93±0.86	57.33±0.71	51.29±0.96	36.03±1.34
APL	54.03±0.92	49.06±0.93	36.06±2.02	55.62±0.92	48.37±0.94	35.02±1.72
S2E	59.37±1.09	43.29±1.94	30.08±3.91	58.92±1.21	42.88±2.16	29.93±4.05
Revision	59.62±0.97	53.26±0.84	35.82±2.06	58.77±0.93	52.72±1.38	37.72±1.75
Reweight	49.59±0.74	39.72±0.57	22.79±1.35	48.87±0.96	36.65±0.90	17.24±1.97
Forward	48.68±0.57	39.78±1.23	27.01±0.89	47.90±0.23	37.89±0.57	21.71±1.53
R-Class2Simi	55.45±0.55	50.38±0.49	35.57±0.75	54.95±0.65	47.56±0.72	34.82±0.58
F-Class2Simi	**60.26±0.18**	**54.85±0.60**	**40.38±0.58**	**59.10±0.13**	**52.99±0.78**	**38.69±2.84**

Table 2. Means and Standard Deviations of Classification Accuracy over 5 trials on text datasets.

NEWS20	Sym-0.2	Sym-0.4	Sym-0.6	Asym-0.2	Asym-0.4	Asym-0.6
Co-teaching	55.32±0.28	51.09±1.06	47.07±0.83	55.29±0.41	53.08±0.26	45.63±0.75
JoCor	52.21±0.70	49.84±0.92	48.83±0.43	55.58±0.27	49.35±0.62	46.21±0.73
PHuber-CE	55.73±0.38	54.33±0.92	45.05±0.49	56.76±0.26	51.15±0.65	41.59±1.05
APL	56.91±0.21	53.12±1.21	43.60±1.28	56.11±0.23	50.93±1.05	43.60±1.28
S2E	57.93±0.37	47.16±1.32	28.53±5.04	54.89±1.92	50.42±1.71	30.67±3.12
Revision	58.06±0.19	52.30±1.73	46.84±1.09	56.41±0.77	53.44±0.83	43.77±1.08
Reweight	53.34±1.08	50.15±1.33	44.73±0.79	53.37±0.66	49.82±0.44	39.46±1.27
Forward	57.30±0.32	53.94±0.42	46.91±1.48	53.58±0.54	49.90±1.44	42.55±3.81
R-Class2Simi	**58.67±0.38**	56.59±0.74	50.48±0.97	58.44±0.66	**55.03±1.55**	**47.75±2.17**
F-Class2Simi	58.27±0.47	**56.70±1.13**	**50.18±0.89**	**58.46±0.68**	54.92±1.66	46.07±3.54

南京航空航天大學

Nanjing University of Aeronautics and Astronautics

Selective-Supervised Contrastive Learning with Noisy Labels

Shikun Li[1,2], Xiaobo Xia[3], Shiming Ge[1,2]*, Tongliang Liu[3]
[1] Institute of Information Engineering, Chinese Academy of Sciences, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, China
[3] Trustworthy Machine Learning Lab, The University of Sydney, Australia
{lishikun,geshiming}@iie.ac.cn, xxia5420@uni.sydney.edu.au, tongliang.liu@sydney.edu.au

CVPR 2022

Overview

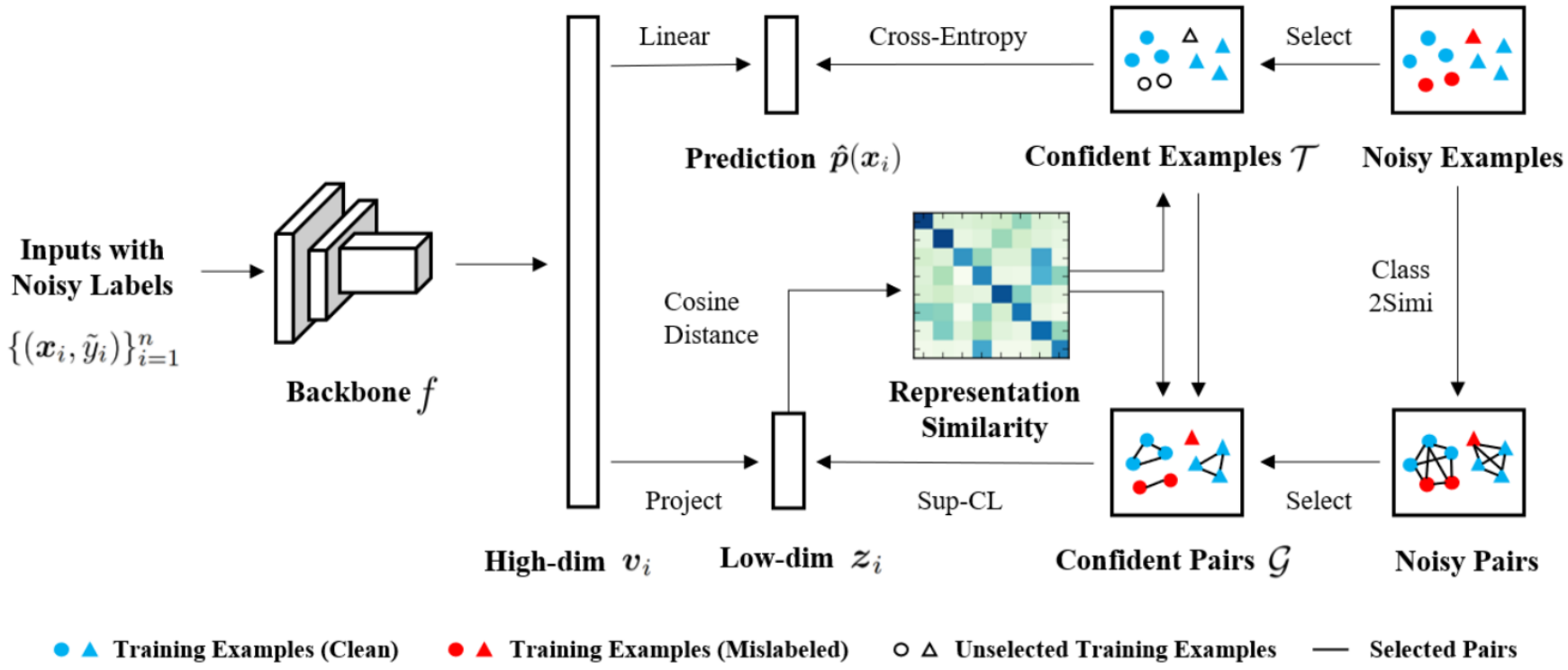


Figure 2. The illustration of the proposed Sel-CL, which progressively selects better confident pairs $\mathcal{G}$ for supervised contrastive learning based on the representation similarity. Without the noise rate prior, confident examples $\mathcal{T}$ are also obtained to help identify the pairs.

Selecting Confident Examples

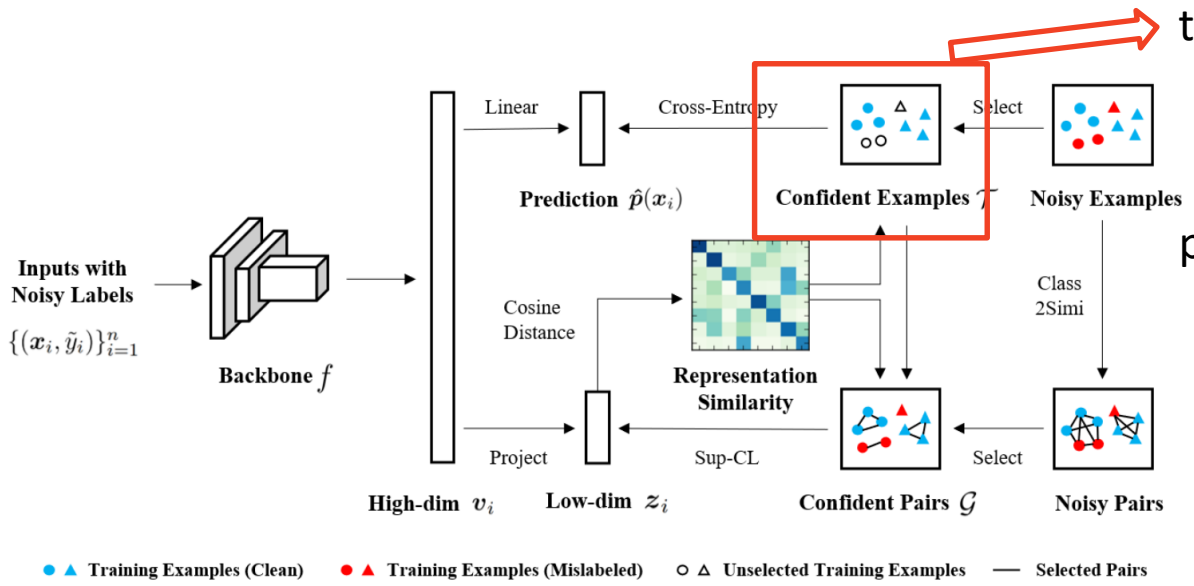


Figure 2. The illustration of the proposed Sel-CL, which progressively selects better confident pairs $\mathcal{G}$ for supervised contrastive learning based on the representation similarity. Without the noise rate prior, confident examples $\mathcal{T}$ are also obtained to help identify the pairs.

given two low-dimensional representations z_i and z_j , calc. the cosine similarity:

$$d(\mathbf{z}_i, \mathbf{z}_j) = \frac{\mathbf{z}_i \mathbf{z}_j^\top}{\| \mathbf{z}_i \| \| \mathbf{z}_j \|}$$

pseudo-labels(Top-K neighbors, K=250):

$$\hat{q}_c(\mathbf{x}_i) = \frac{1}{K} \sum_{\substack{k=1 \\ \mathbf{x}_k \in \mathcal{N}_i}}^{K} \mathbb{I}\,[\hat{y}_k = c], c \in [C]$$

$\mathcal{T}_c$ confident examples belonging to the c-th class:

$$\mathcal{T}_c = \{(\mathbf{x}_i, \tilde{y}_i) \mid \ell(\hat{\mathbf{q}}(\mathbf{x}_i), \tilde{y}_i) < \gamma_c, i \in [n]\}, c \in [C]$$

the confident example set including all classes

$$\mathcal{T} = \bigcup_{c=1}^{C} \mathcal{T}_c$$

Selecting Confident Pairs

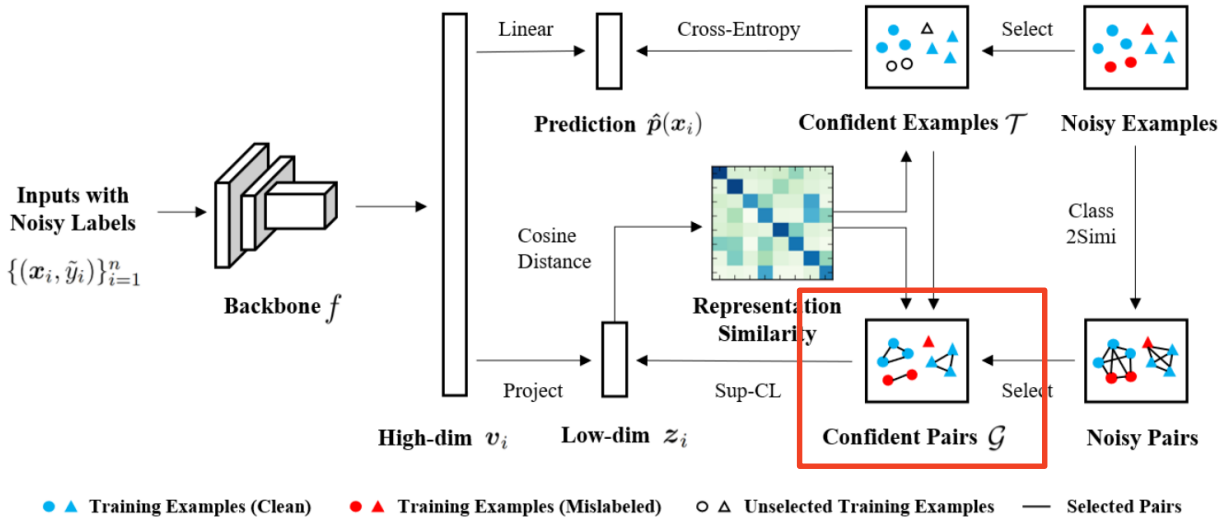


Figure 2. The illustration of the proposed Sel-CL, which progressively selects better confident pairs $\mathcal{G}$ for supervised contrastive learning based on the representation similarity. Without the noise rate prior, confident examples $\mathcal{T}$ are also obtained to help identify the pairs.

confident examples -> confident pairs:

$$\mathcal{G}' = \{P_{ij} \mid \tilde{y}_i = \tilde{y}_j, (\mathbf{x}_i, \tilde{y}_i), (\mathbf{x}_j, \tilde{y}_j) \in \mathcal{T}\}$$

Noisy positive pairs - > confident pairs:

$$\mathcal{G}'' = \{P_{ij} \mid \tilde{s}_{ij} = 1, d(\mathbf{z}_i, \mathbf{z}_j) > \gamma\}$$

The confident pair set:

$$\mathcal{G} = \mathcal{G}' \cup \mathcal{G}''$$

Representation Learning with Selected Pairs

Contrastive learning(Sup-CL):

$$\mathcal{L} = \sum_{i \in I} \mathcal{L}_i(\mathbf{z}_i) = \sum_{i \in I} \frac{-1}{|\mathcal{G}(i)|} \sum_{g \in \mathcal{G}(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_g / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)}$$

Mixup loss:

$$\mathcal{L}_i^{\mathrm{MIX}}(\mathbf{z}_i) = \lambda \mathcal{L}_a(\mathbf{z}_i) + (1-\lambda)\mathcal{L}_b(\mathbf{z}_i)$$

Classification learning:

$$\mathcal{L}^{\mathrm{CLS}} = \sum_{(\mathbf{x}_i, \tilde{y}_i) \in \mathcal{T}} \mathcal{L}_i^{cls}(\mathbf{x}_i) = \sum_{(\mathbf{x}_i, \tilde{y}_i) \in \mathcal{T}} \ell(\hat{\mathbf{p}}(\mathbf{x}_i), \tilde{y}_i)$$

Similarity loss:

$$\mathcal{L}^{\mathrm{SIM}} = \sum_{i \in I} \sum_{j \in A(i)} \ell\left(\hat{\mathbf{p}}(\mathbf{x}_i)\hat{\mathbf{p}}(\mathbf{x}_j), \mathbb{I}[P_{i'j'} \in \mathcal{G}]\right)$$

$$\mathcal{L}^{\mathrm{ALL}} = \mathcal{L}^{\mathrm{MIX}} + \lambda_c \mathcal{L}^{\mathrm{CLS}} + \lambda_s \mathcal{L}^{\mathrm{SIM}}$$

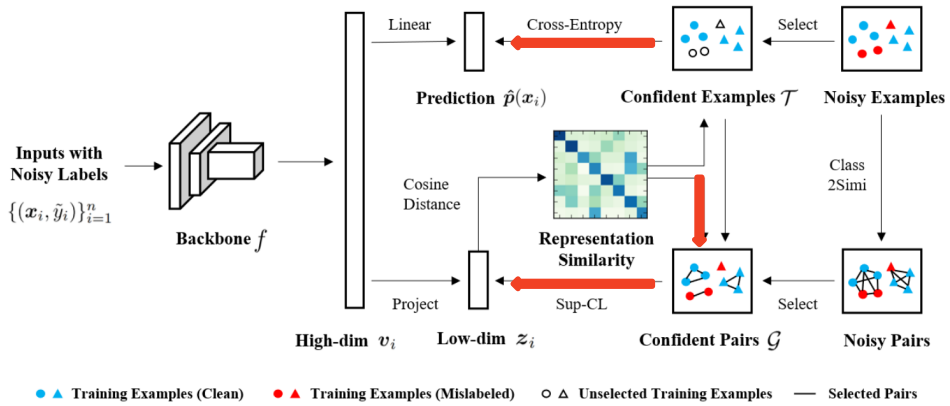


Figure 2. The illustration of the proposed Sel-CL, which progressively selects better confident pairs $\mathcal{G}$ for supervised contrastive learning based on the representation similarity. Without the noise rate prior, confident examples $\mathcal{T}$ are also obtained to help identify the pairs.

Experiment

Table 3. Comparison with state-of-the-art methods in the test accuracy (%) on CIFAR-10 and CIFAR-100. The best results are in **bold**.

Dataset	CIFAR-10								CIFAR-100							
Methods/Noise rate	Symmetric				Asymmetric				Symmetric				Asymmetric			
	20%	50%	80%	90%	10%	20%	30%	40%	20%	50%	80%	90%	10%	20%	30%	40%
Cross-Entropy	82.7	57.9	26.1	16.8	88.8	86.1	81.7	76.0	61.8	37.3	8.8	3.5	68.1	63.6	53.3	44.5
Mixup [61]	92.3	77.6	46.7	43.9	93.3	88.0	83.3	77.7	66.0	46.6	17.6	8.1	72.4	65.1	57.6	48.1
Forward [43]	83.1	59.4	26.2	18.8	90.4	86.7	81.9	76.7	61.4	37.3	9.0	3.4	68.7	63.2	54.4	45.3
GCE [64]	86.6	81.9	54.6	21.2	89.5	85.6	80.6	76.0	59.2	47.8	15.8	7.2	68.0	58.6	51.4	42.9
P-correction [59]	92.0	88.7	76.5	58.2	93.1	92.9	92.6	91.6	68.1	56.4	20.7	8.8	76.1	68.9	59.3	48.3
M-correction [1]	93.8	91.9	86.6	68.7	89.6	91.8	92.2	91.2	73.4	65.4	47.6	20.5	67.1	64.5	58.6	47.4
DivideMix [30]	95.0	93.7	**92.4**	74.2	93.8	93.2	92.5	91.4	74.8	72.1	57.6	29.2	69.5	69.2	68.3	51.0
ELR [37]	93.8	92.6	88.0	63.3	94.4	93.3	91.5	85.3	74.5	70.2	45.2	20.5	75.8	74.8	73.6	70.0
GCE (Uns-CL init.) [13]	90.0	89.3	73.9	36.5	91.1	87.3	82.2	78.1	68.1	53.3	22.1	8.9	70.2	60.2	52.6	44.1
ELR (Uns-CL init.)	94.4	93.0	88.3	**86.2**	95.0	94.7	94.4	93.3	76.2	71.9	57.9	40.8	77.2	75.5	74.3	70.4
MOIT+ [41]	94.1	91.8	81.1	74.7	94.2	94.3	94.3	93.3	75.9	70.6	47.6	41.8	77.4	76.4	75.1	74.0
Sel-CL+	**95.5**	**93.9**	89.2	81.9	**95.6**	**95.2**	**94.5**	**93.4**	**76.5**	**72.4**	**59.6**	**48.8**	**78.7**	**77.5**	**76.4**	**74.2**

Ablation Study and Discussions

Table 6. Ablation study for Sel-CL and Sel-CL+ on CIFAR-100. The best results are in **bold**.

Methods	Sym. 20%	Asym. 40%
Sel-CL w/o Mixup Data Aug.	70.3/70.6	64.2/66.2
Sel-CL w/o MOCO Trick	73.3/74.1	69.2/71.5
Sel-CL w/o Selection	67.2/68.9	49.9/68.7
Sel-CL w/o Classfier Learning	— /69.9	— /70.2
Sel-CL w/o $\mathcal{L}^{SIM}$	74.5/74.9	71.8/72.5
Sel-CL	**74.9/75.4**	**72.0/72.7**
Sel-CL+ w/ Strong Data Aug.	74.5	72.7
Sel-CL+ w/o Retraining Cls.	76.4	73.4
Sel-CL+	**76.5**	**74.2**

Ablation study

Table 7. Comparison with different warm-up methods in the test accuracy (%) of Sel-CL+.

Dataset	CIFAR-10			CIFAR-100		
Noise type	Sym.		Asym.	Sym.		Asym.
Noise rate	20%	90%	40%	20%	90%	40%
Uns-CL [7]	95.5	81.9	93.4	76.5	48.8	74.2
Sup-CL [27]	95.5	81.6	93.4	76.8	51.4	74.5

Discussions on warm-up methods.

Table 8. Comparison with using different fine-tuning methods in the test accuracy (%). The best results are in **bold**.

Dataset	CIFAR-10		CIFAR-100	
Noise type	Sym.	Asym.	Sym.	Asym.
Noise rate	20%	40%	20%	40%
DivideMix [30]	95.7	92.1	76.9	53.8
ELR+ [37]	94.6	93.0	77.5	72.2
DivideMix (Uns-CL init.) [65]	96.2	90.8	78.3	52.9
ELR+ (Uns-CL init.) [65]	94.8	94.3	77.7	72.3
DivideMix (Sel-CL init.)	**96.3**	91.6	**78.7**	55.2
ELR+ (Sel-CL init.)	95.2	**94.6**	77.7	**72.9**

Discussions on fine-tuning methods

THANKS