Prototype-Anchored Learning For Learning With Imperfect Annotations

Xiong Zhou[1], Xianming Liu[1 2], Deming Zhai[1], Junjun Jiang[1 2], Xin Gao[3 2], Xiangyang Ji[4]

[1] *Harbin Institute of Technology*

[2] *Peng Cheng Laboratory*

[3] *King Abdullah University of Science and Technology*

[4] *Tsinghua University*

ICML International Conference On Machine Learning

ICML | 2022
Thirty-ninth International Conference on Machine Learning

Motivation and Our Contributions

- Motivation:
  - High-quality annotated data are usually difficult or expensive to obtain.
  - The resulting labels may be class-imbalanced, noisy or human biased.
  - It is challenging to learn robust and unbiased models from imperfectly annotated datasets.

Motivation and Our Contributions

- Motivation:
  - High-quality annotated data are usually difficult or expensive to obtain.
  - The resulting labels may be class-imbalanced, noisy or human biased.
  - It is challenging to learn robust and unbiased models from imperfectly annotated datasets.
- Our Contributions:
  - A theoretically sound, simple yet effective scheme—Prototype-Anchored Learning (PAL).
  - For class-imbalanced learning, PAL can implicitly guarantee balanced representations.
  - For learning with noisy labels, we extend the classical symmetric condition and reveal that PAL can lead to a tighter bound.

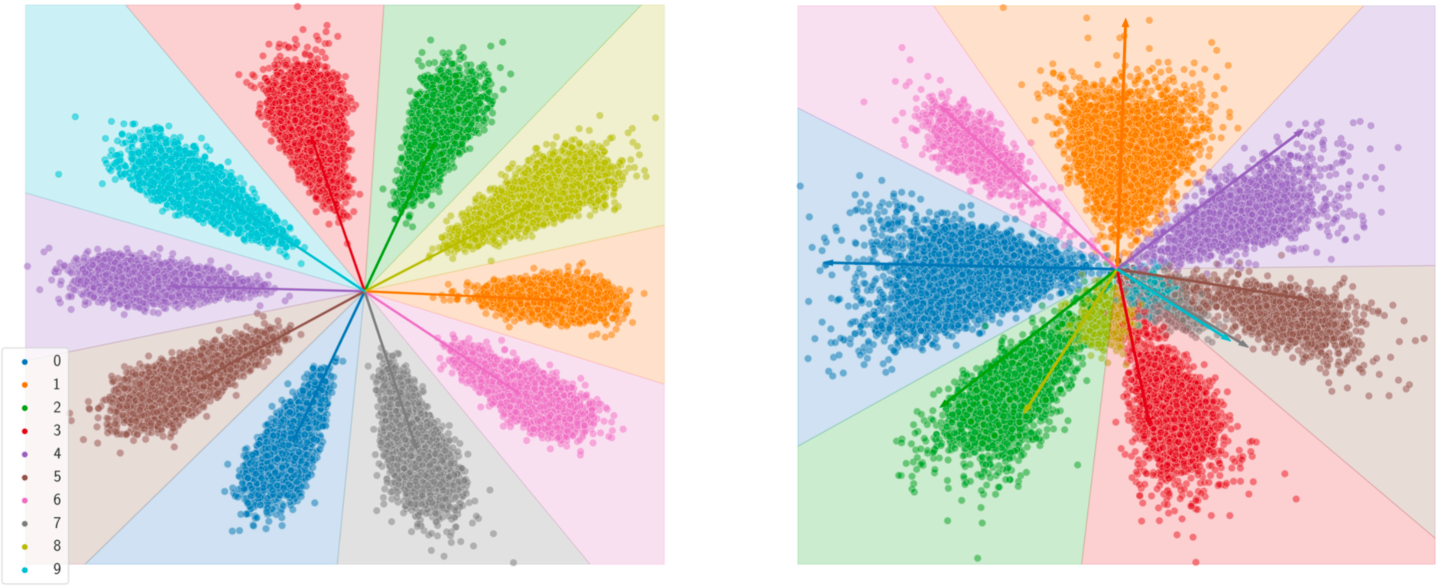
ICML
International Conference
On Machine Learning

Preliminaries

The softmax loss: For a labeled dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, the softmax loss for a k-classification problem is formulated as

$$L_i = -\log \frac{\exp(\boldsymbol{w}_{y_i}^{\top} \boldsymbol{z}_i)}{\sum_{j=1}^{k} \exp(\boldsymbol{w}_j^{\top} \boldsymbol{z}_i)},$$

where $z_i = \phi_\Theta(x_i) \in \mathbb{R}^d$ (usually $\mathrm{k} \leq d + 1$) is the learned feature representation vector , ϕ_Θ denotes the feature extraction sub-network, $W = (w_1, \dots, w_k) \in \mathbb{R}^{d \times k}$ denotes the linear classifier which is implemented with a linear layer.



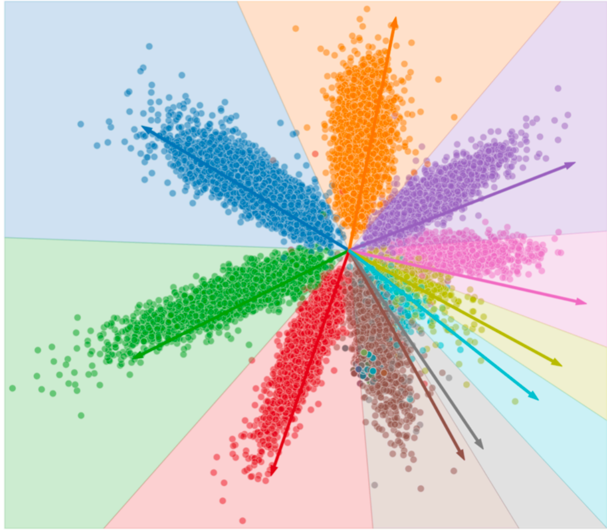
(a) (b)

Figure 1. Visualization on MNIST (a) and long-tailed MNIST (b) by the Softmax loss. (a) denotes the class-balanced case by CE, where features and prototypes are optimized to be perfectly balanced. (b) denotes the class-imbalanced case by CE, where the majority classes (“0-3”) occupy most of the feature space, the representations of minority classes (“7-9”) are narrow, and the majority classes have larger norms and angular distance from other prototypes, and the reverse on the minority classes.

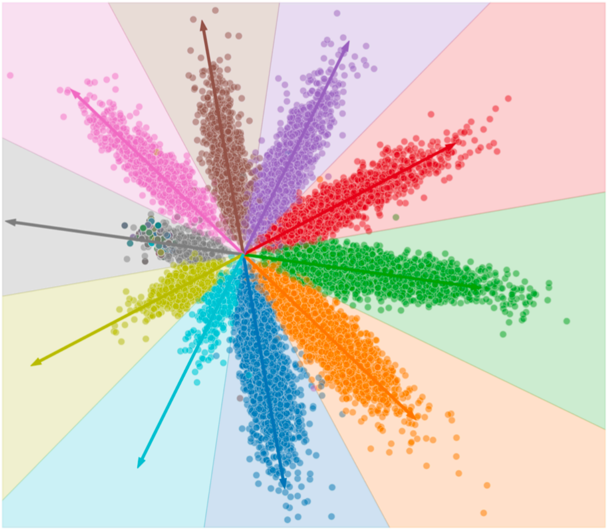
Preliminaries

Margin-based loss: By requiring features and prototypes on the unit sphere, margin-based losses[1] introduce a margin to obtain strong discriminativeness:

$$L_{\boldsymbol{\alpha}} = -\log \frac{\exp(s\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z} + \alpha_y)}{\exp(s\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z} + \alpha_y) + \sum_{j \neq y} \exp(s\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{z})},$$



(a) Normalization on features and prototypes



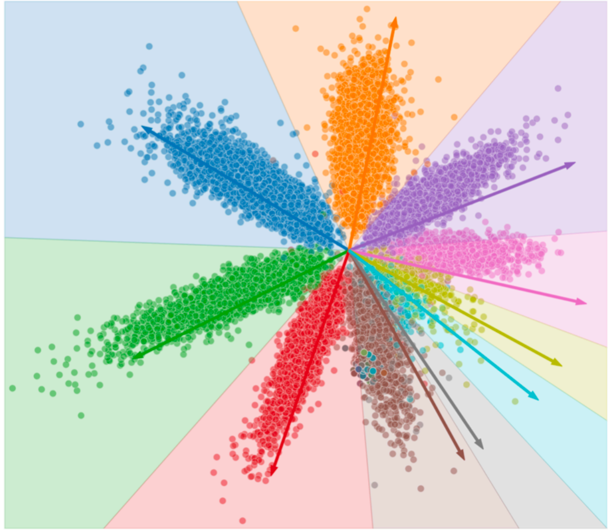
(b) PAL–based

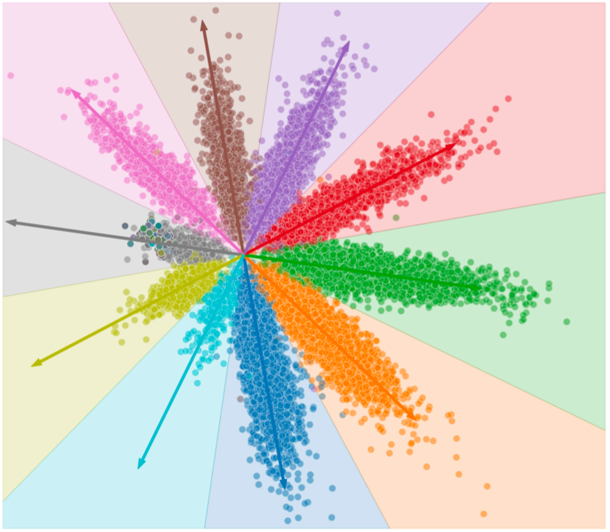
[1] Cao et. al. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss[J]. Advances in Neural Information Processing Systems, 2019, 32: 1567-1578.

Preliminaries

Margin-based loss: By requiring features and prototypes on the unit sphere, margin-based losses[1] introduce a margin to obtain strong discriminativeness:

$$L_{\alpha} = -\log \frac{\exp(s\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z} + \alpha_y)}{\exp(s\boldsymbol{w}_y^{\mathrm{T}}\boldsymbol{z} + \alpha_y) + \sum_{j \neq y} \exp(s\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{z})},$$





which coincides with the goal of tightening a class-balanced generalization error bound

$$\mathbb{P}_{(\boldsymbol{x},y)}[f(\boldsymbol{x})_y < \max_{l \neq y} f(\boldsymbol{x})_l]$$
$$\leq \frac{1}{k}\sum_{j=1}^{k}\left(\hat{L}_{\gamma_j,j}[f] + \frac{4}{\gamma_j}\hat{\mathfrak{R}}_j(\mathcal{F}) + \varepsilon_j(\gamma_j)\right)$$

where γ_j is the sample margin for class j.

Sample Margin: The sample margin of (x, y) is defined as

$$\gamma(\boldsymbol{x}, y) = f(\boldsymbol{x})_y - \max_{j \neq y} f(\boldsymbol{x})_j = \boldsymbol{w}_y^{\top}\boldsymbol{z} - \max_{j \neq y} \boldsymbol{w}_j^{\top}\boldsymbol{z},$$

the sample margin for class j is $\gamma_j = \min_{i \in S_j} \gamma(x_i, y_i)$, and the minimal sample margin is $\gamma_{\min} = \min\{\gamma_1, \dots, \gamma_k\}$.

We can maximize $\gamma_{\min}$ to tighten error bound for each class!

[1] Cao et. al. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss[J]. Advances in Neural Information Processing Systems, 2019, 32: 1567-1578.

The optimality of maximizing $\gamma_{\min}$

International Conference On Machine Learning

Lemma 1. [The Optimality Condition of prototypes to Maximize $\gamma_{\min}$] If $w_1, \ldots, w_k, z_1, \ldots, z_N \in \mathbb{S}^{d-1}$ $(2 \leq k \leq d+1)$, then the maximum of the minimal sample margin $\gamma_{\min}$ is $\frac{k}{k-1}$, which is uniquely obtained if $z_i = w_{y_i}, \forall i$, and $w_i^T w_j = -\frac{1}{k-1}, \forall i \neq j$.

Theorem 2. For balanced datasets (i.e., each class has the same number of samples), if $w_1, \ldots, w_k, z_1, \ldots, z_N \in \mathbb{S}^{d-1}$ $(2 \leq k \leq d+1)$, then learning with L_α that has the same per-class margins (i.e., $\alpha_j = \alpha, \forall j \in [k]$) will deduce $z_i = w_{y_i}, \forall i$, and $w_i^T w_j = -\frac{1}{k-1}, \forall i \neq j$.

[1] Cao et. al. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss[J]. Advances in Neural Information Processing Systems, 2019, 32: 1567-1578.

The optimality of maximizing $\gamma_{\min}$

ICML
International Conference
On Machine Learning

Lemma 1. [The Optimality Condition of prototypes to Maximize $\gamma_{\min}$] If $w_1, \dots, w_k, z_1, \dots, z_N \in \mathbb{S}^{d-1}$ $(2 \leq k \leq d+1)$, then the maximum of the minimal sample margin $\gamma_{\min}$ is $\frac{k}{k-1}$, which is uniquely obtained if $z_i = w_{y_i}, \forall i$, and $w_i^T w_j = -\frac{1}{k-1}, \forall i \neq j$.

Theorem 2. For balanced datasets (i.e., each class has the same number of samples), if $w_1, \dots, w_k, z_1, \dots, z_N \in \mathbb{S}^{d-1}$ $(2 \leq k \leq d+1)$, then learning with L_α that has the same per-class margins (i.e., $\alpha_j = \alpha, \forall j \in [k]$) will deduce $z_i = w_{y_i}, \forall i$, and $w_i^T w_j = -\frac{1}{k-1}, \forall i \neq j$.

Theorem 3. Under class-imbalanced data distribution (where we have different per-class margins), LDAM[1] is not classification-calibrated.

[1] Cao et. al. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss[J]. Advances in Neural Information Processing Systems, 2019, 32: 1567-1578.

Prototype-Anchored Learning (PAL)

- Lemma 1 provides the optimality condition of prototypes to maximizing the minimal sample margin, that is,

$$w_i^T w_j = -\frac{1}{k-1}, \forall i \neq j.$$

- We then propose to initialize a group of prototypes that satisfying the above equation, and this method is called as *prototype-anchored learning* (PAL).

- The desired prototypes can be easily obtained according Theorem 2.

```
def generate_weight(n_classes, n_hiddens, use_relu=False):
    n_samples = n_classes
    scale = 5
    Z = torch.randn(n_samples, n_hiddens).cuda()
    Z.requires_grad = True
    W = torch.randn(n_classes, n_hiddens).cuda()
    W.requires_grad = True
    nn.init.kaiming_normal_(W)

    optimizer = SGD([Z, W], lr=0.1, momentum=0.9, weight_decay=1e-4)
    scheduler = CosineAnnealingLR(optimizer, T_max=20000, eta_min=0)

    criterion = nn.CrossEntropyLoss()
    for i in range(epochs):
        if use_relu:
            z = F.relu(Z)
        else:
            z = Z
        w = W
        L2_z = F.normalize(z, dim=1)
        L2_w = F.normalize(w, dim=1)
        out = F.linear(L2_z, L2_w)
        loss = criterion(out * scale, labels)
        optimizer.zero_grad()
        loss.backward()
        optimizer.step()
        scheduler.step()
    return W
```

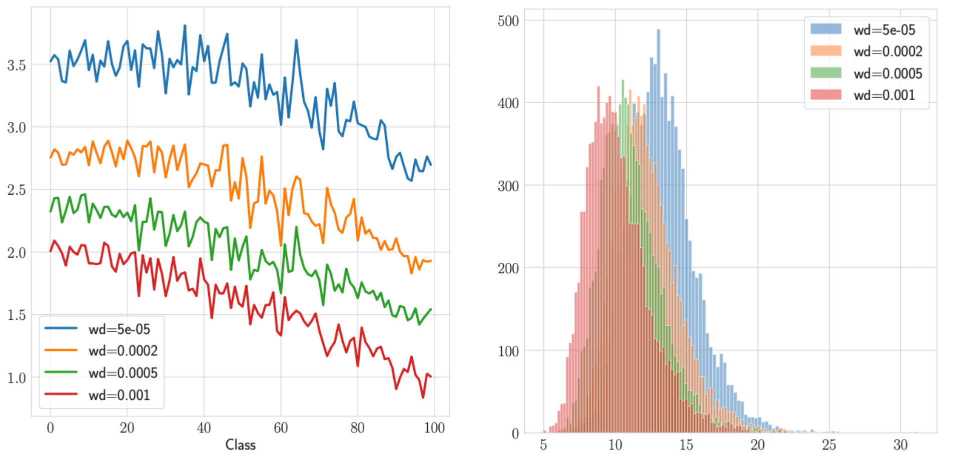
PAL for Class-imbalanced Learning

Theorem 4. For imbalanced or balanced datasets, if $w_1, \dots, w_k, \ z_1, \dots, z_N \in \mathbb{S}^{d-1}$ $(2 \leq k \leq d+1)$, where $w_1, \dots, w_k$ are anchored and satisfy that $w_i^T w_j = -\frac{1}{k-1}$, $\forall i \neq j$, then learning with L_α will deduce deduce $z_i = w_{y_i}, \forall i$, and the minimal sample margin $\gamma_{\min}$ will be $\frac{k}{k-1}$.

PAL for Class-imbalanced Learning

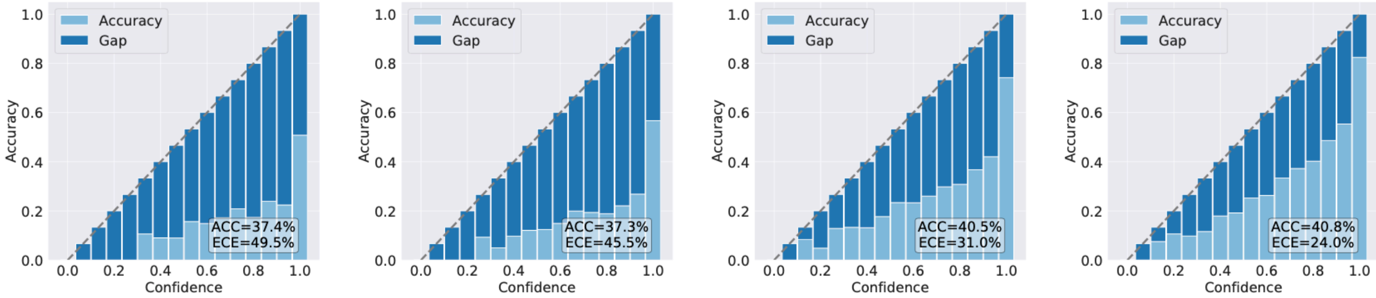
Theorem 4. For imbalanced or balanced datasets, if $w_1, \dots, w_k,\ z_1, \dots, z_N \in \mathbb{S}^{d-1}$ $(2 \le k \le d+1)$, where $w_1, \dots, w_k$ are anchored and satisfy that $w_i^T w_j = -\frac{1}{k-1}$, $\forall i \neq j$, then learning with L_α will deduce deduce $z_i = w_{y_i}, \forall i$, and the minimal sample margin $\gamma_{\min}$ will be $\frac{k}{k-1}$.

Theorem 5. For imbalanced or balanced datasets, if $||z_i||_2 \le B, \forall i \in [N]$), and the prototypes $w_1, \dots, w_k$ are anchored to satisfy $w_i^T w_j = -\frac{1}{k-1}$, $\forall i \neq j$, then learning with the softmax loss will deduce deduce $\frac{w_{y_i}^T z_i}{||w_{y_i}||_2 ||z_i||_2}$=1, $\forall i$, and obtain the maximum of the minimal sample margin $\gamma_{\min}$.



(a) prototype norms (b) histogram of feature norms

Figure 2. Illustration of prototypes norms and feature norms by CE trained on CIFAR-100-LT with imbalance ratio 100 under different weight decays. As can be seen, the larger weight decay usually leads to smaller prototype norms and feature norms.



(a) CE with wd=0.0 (b) CE with wd=5e-5 (c) CE with wd=5e-4 (d) CE with wd=1e-3

Figure 3. Reliability diagrams of ResNet-32 trained by CE on CIFAR-100-LT with imbalance ratio 100 under different weight decays (wds). As can be seen, an appropriate larger weight decay can improve both accuracy and confidence.

PAL for Noise-Tolerant Learning

The most popular family of loss functions is **symmetric losses**, which satisfies

$$\sum_{i=1}^{k} L(f(\boldsymbol{x}), i) = \mathrm{C}, \forall x \in \mathcal{X}, \forall f,$$

where C is a constant. This symmetric condition theoretically guarantees the noise tolerance by risk minimization under symmetric label noise, i.e., the global minimizer of the noisy L-risk $R_L^{\eta}(f) = \mathbb{E}_{\boldsymbol{x},\tilde{y}}[L(f(x), \tilde{y})] = \mathbb{E}_{\boldsymbol{x},y}\left[(1-\eta_x)L(f(x), y) + \sum_{i \neq y} \eta_{\boldsymbol{x},i} L(f(x), i)\right]$ also minimizes the L-risk $R_L(f) = \mathbb{E}_{\mathcal{D}} L(f(x), y)$, where $\eta_{\boldsymbol{x},i}$ denotes noise rates.

Negative-signed Sample Logit Loss (NSL): $L_{NSL}(f(x), i) = -f(x)_i = -w_i^T \phi_\Theta(x)$.

Proposition 6. If the prototypes $w_1, \dots, w_k \in \mathbb{S}^{d-1}$ are anchored to satisfy $w_i^T w_j = -\frac{1}{k-1}, \forall i \neq j$, then $L_{NSL}(f(x), i) = -w_i^T \phi_\Theta(x)$ is symmetric. More specifically, we have. $\sum_{i=1}^{k} L_{NSL}(f(x), i) = 0$, and learning with L_{NSL} will lead to the maximum of $\gamma_{\min}$ under symmetric label noise.

PAL for Noise-Tolerant Learning

Theorem 7. In a multi-class classification problem, given $w_1, \ldots, w_k$, if $z = \phi_\Theta(x)$ is norm-bounded by B, i.e., $||z||_2 \leq B$, then for any loss $L(z, i)$ satisfying $L_W(z) = \sum_{i=1}^{k} L(W^T z, i)$ is λ-Lipschitz, we have the following risk bound under symmetric label noise with $\eta < \frac{k-1}{k}$:

$$R_L(\hat{f}) - R_L(f^*) \leq \frac{2\eta\lambda B}{(1-\eta)k - 1},$$

where $\hat{f}$ and f^* denote a global minimizer of $R_L^\eta(f)$ and $R_L(f)$, respectively.

Proposition 8. In a multi-class classification problem, let $w_1, \ldots, w_k \in \mathbb{S}^{d-1}$ $(2 \leq k \leq d+1)$ satisfy $w_i^T w_j = -\frac{1}{k-1}, \forall i \neq j$, if $z = \phi_\Theta(x)$ is norm-bounded by B, i.e., $||z||_2 \leq B$, then we have the following risk bound for the CE loss under symmetric label noise with $\eta < \frac{k-1}{k}$:

$$R_L(\hat{f}) - R_L(f^*) \leq \frac{2c\eta k(1-t)B}{k - 1 + t(k-1)^2},$$

wherec $= \frac{k-1}{(1-\eta)k-1}$, $t = \exp(-\frac{kB}{k-1})$, $\hat{f}$ and f^* denote a global minimizer of $R_L^\eta(f)$ and $R_L(f)$, respectively.

Experimental Results

Table 1. Validation accuracy (%) on ImageNet-LT. The results with positive gains are **boldfaced** and the best one is underlined.

Method	Many	Medium	Few	All
CE	66.8	36.9	7.1	43.6
FL	64.3	37.1	8.2	43.7
OLTR	51.0	40.8	20.8	41.9
Causal Norm	65.2	47.7	29.8	52.0
Balanced Softmax	63.6	48.4	32.9	52.1
LADE	65.1	48.9	33.4	53.0
cRT+mixup	63.9	49.1	30.2	51.7
LWS+mixup	62.9	49.8	31.6	52.0
MiSLAS	61.7	51.3	35.8	52.7
CE+PAL	**69.0**	**42.5**	**11.0**	**47.6**
MiSLAS+PAL	**64.0**	**51.6**	32.4	**53.3**

Table 5. Top-1 validation accuracies (%) on mini-WebVision.

Method	CE	FL	NCE+RCE	**NSL**	**CE+PAL**	**CE+FNPAL**
Acc	62.60	63.80	66.32	**69.56**	**68.92**	**69.69**

Table 3. Validation accuracies (%) of different methods on benchmark datasets with clean or symmetric label noise ($\eta \in [0.2, 0.4, 0.6, 0.8]$). The results (mean ± std) are reported over 3 random runs. The results with positive gains are **boldfaced** and the best one is underlined.

Dataset	Method	Clean ($\eta = 0.0$)	Symmetric Noise Rate (η) 0.2	0.4	0.6	0.8
MNIST	CE	99.17 ± 0.04	91.40 ± 0.11	74.36 ± 0.29	49.32 ± 0.70	22.32 ± 0.15
	FL	99.16 ± 0.02	91.49 ± 0.20	75.28 ± 0.10	50.25 ± 0.70	22.68 ± 0.14
	GCE	99.15 ± 0.02	98.90 ± 0.03	96.81 ± 0.23	81.39 ± 0.64	33.07 ± 0.31
	SCE	99.28 ± 0.07	98.91 ± 0.12	97.60 ± 0.22	88.00 ± 0.50	47.32 ± 0.99
	NCE+MAE	99.42 ± 0.02	99.18 ± 0.08	98.47 ± 0.21	95.52 ± 0.04	73.05 ± 0.59
	NCE+RCE	99.40 ± 0.04	99.24 ± 0.01	98.44 ± 0.11	95.77 ± 0.09	74.80 ± 0.28
	NFL+RCE	99.37 ± 0.01	99.16 ± 0.03	98.55 ± 0.05	95.62 ± 0.24	74.67 ± 0.97
	NSL	99.24 ± 0.03	98.99 ± 0.03	98.58 ± 0.11	95.99 ± 0.24	59.77 ± 1.98
	CE+FNPAL	99.24 ± 0.05	**99.05 ± 0.04**	**98.66 ± 0.04**	**97.62 ± 0.15**	**79.23 ± 0.87**
	SCE+FNPAL	99.27 ± 0.04	**99.06 ± 0.05**	**98.76 ± 0.09**	**97.94 ± 0.07**	**88.56 ± 1.07**
	NCE+RCE+FNPAL	99.29 ± 0.04	99.04 ± 0.07	98.11 ± 0.09	94.84 ± 0.08	**79.70 ± 1.06**
	NFL+RCE+FNPAL	99.29 ± 0.06	99.02 ± 0.05	98.32 ± 0.14	95.38 ± 0.11	**76.06 ± 0.58**
CIFAR10	CE	90.36 ± 0.25	74.78 ± 0.68	57.95 ± 0.12	38.21 ± 0.12	18.89 ± 0.43
	FL	89.69 ± 0.25	74.19 ± 0.23	57.35 ± 0.27	38.11 ± 0.76	19.39 ± 0.44
	GCE	89.37 ± 0.29	87.05 ± 0.21	82.43 ± 0.10	68.05 ± 0.07	25.21 ± 0.28
	SCE	91.24 ± 0.19	87.34 ± 0.01	79.84 ± 0.43	61.09 ± 0.19	27.19 ± 0.34
	NCE+MAE	89.02 ± 0.09	87.06 ± 0.17	83.92 ± 0.16	76.47 ± 0.25	45.01 ± 0.31
	NCE+RCE	91.12 ± 0.14	89.21 ± 0.00	86.03 ± 0.14	80.04 ± 0.26	51.67 ± 1.38
	NFL+RCE	91.03 ± 0.15	89.10 ± 0.16	86.20 ± 0.19	79.58 ± 0.08	50.03 ± 2.78
	NSL	88.07 ± 0.12	86.46 ± 0.02	83.27 ± 0.13	76.17 ± 0.40	46.74 ± 0.72
	CE+FNPAL	90.69 ± 0.11	**86.34 ± 0.37**	**81.30 ± 0.29**	**72.77 ± 0.41**	**51.46 ± 1.10**
	SCE+FNPAL	91.11 ± 0.13	87.30 ± 0.06	**82.68 ± 0.22**	**73.49 ± 0.42**	**51.99 ± 1.10**
	NCE+RCE+FNPAL	90.88 ± 0.10	**89.34 ± 0.15**	**86.65 ± 0.21**	**80.28 ± 0.07**	**57.21 ± 0.22**
	NFL+RCE+FNPAL	91.16 ± 0.25	**89.49 ± 0.32**	**86.66 ± 0.08**	**80.33 ± 0.15**	**56.23 ± 0.15**
CIFAR100	CE	70.41 ± 1.17	55.64 ± 0.17	40.39 ± 0.46	22.00 ± 1.23	7.37 ± 0.16
	FL	70.56 ± 0.59	56.02 ± 0.80	40.41 ± 0.39	22.11 ± 0.30	7.70 ± 0.20
	GCE	63.06 ± 1.00	62.15 ± 0.66	57.11 ± 1.43	45.99 ± 1.00	18.32 ± 0.36
	SCE	70.41 ± 0.63	55.05 ± 0.68	39.60 ± 0.14	21.53 ± 0.72	7.82 ± 0.30
	NCE+MAE	67.16 ± 0.13	52.34 ± 0.12	35.81 ± 0.42	19.29 ± 0.29	7.31 ± 0.23
	NCE+RCE	68.09 ± 0.26	64.32 ± 0.40	58.11 ± 0.63	45.94 ± 1.31	25.22 ± 0.08
	NFL+RCE	67.58 ± 0.39	64.48 ± 0.50	57.86 ± 0.12	46.74 ± 0.59	24.55 ± 0.47
	NSL	70.08 ± 0.19	65.30 ± 0.36	56.77 ± 0.52	41.21 ± 1.01	12.16 ± 0.96
	CE+FNPAL	71.69 ± 0.27	**65.38 ± 0.17**	**57.24 ± 0.36**	**41.35 ± 0.19**	**12.12 ± 0.88**
	SCE+FNPAL	70.87 ± 0.45	**65.30 ± 0.15**	**55.10 ± 0.45**	**39.73 ± 0.04**	**11.70 ± 0.53**
	NCE+RCE+FNPAL	69.29 ± 0.32	**65.53 ± 0.30**	**60.53 ± 0.27**	**49.73 ± 0.64**	24.54 ± 0.28
	NFL+RCE+FNPAL	69.53 ± 0.05	**65.94 ± 0.32**	**60.89 ± 0.60**	**50.10 ± 0.40**	24.15 ± 1.06

ICML International Conference On Machine Learning

Thanks for your attention!

Any question? Please contact us!

Xianming Liu: : csxm@hit.edu.cn
Xiong Zhou: cszx@hit.edu.cn