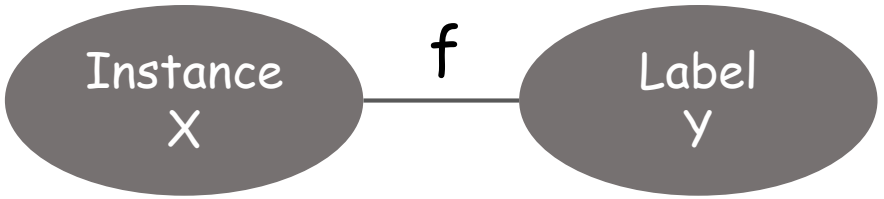
PARNEC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

Revisiting Consistency Regularization for Deep Partial Label Learning

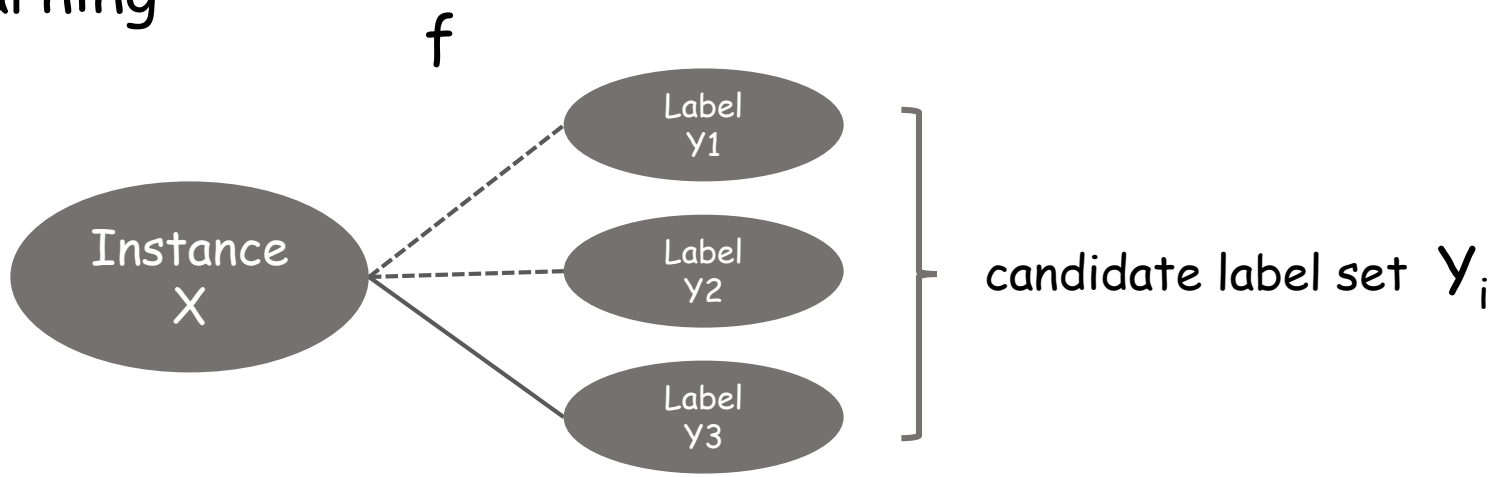
ICML 2022

Partial Label Learning

- Supervised Learning



- Partial Label Learning



Ground-truth ?

Partial Label Learning

A dog image $\boldsymbol{x}_i$ with

$Y_i = \{\text{Husky}, \underline{\text{Malamute}}, \text{Samoyed}\}$

An input image with three candidate labels, where the ground-truth is Malamute

Partial Label Learning

- The key challenge of PLL is to identify the ground-truth label from the candidate label set.

- Learning a classifier from the candidate set Y_i

$$\mathcal{L}_{\text{cls}}(f; \boldsymbol{x}_i, Y_i) = \sum_{j=1}^{C} -s_{i,j} \log(f^j(\boldsymbol{x}_i)) \text{ s.t. } \sum_{j \in Y_i} s_{i,j} = 1 \text{ and } s_{i,j} = 0, \forall j \notin Y_i, \quad (1)$$

- **Disambiguation:**

  Recover the ground-truth confidence s from Y_i

Method

- initialization

$$p_k = \begin{cases} \frac{1}{|\mathcal{S}|} & \text{if } k \in \mathcal{S}, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

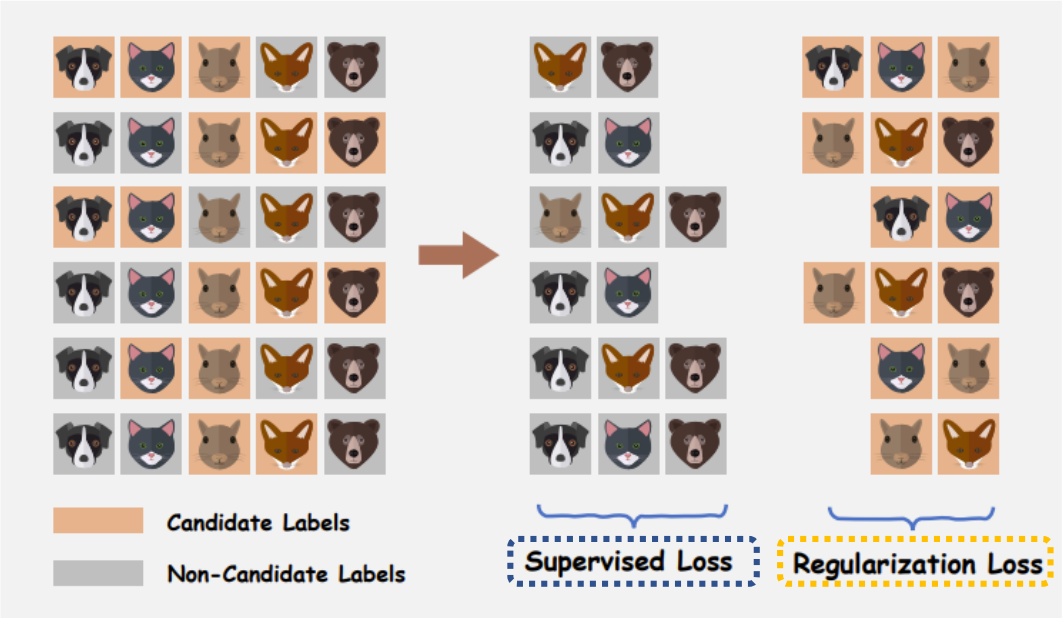


Figure 1. The label matrix is decoupled into two complementary parts including candidate labels and non-candidate labels, on which the supervised loss and regularization loss could be calculated respectively.

$$\mathcal{L}_{\text{Sup}}(\boldsymbol{x}, \mathcal{S}) = -\sum\nolimits_{k \notin \mathcal{S}} \log(1 - g_k(\boldsymbol{x})). \tag{2}$$

- The complements of the candidate labels would never be the ground-truth labels

- Objective function

$$\mathcal{L}(\boldsymbol{x}, \mathcal{S}) = \mathcal{L}_{\text{Sup}}(\boldsymbol{x}, \mathcal{S}) + \lambda \boldsymbol{\Psi}(\boldsymbol{x}, \mathcal{S}), \tag{3}$$

Consistency Regularization

- Minimizing the divergence of each output pair given a set of random augmented instances in

$$\mathcal{A}(\boldsymbol{x}) = \{\mathrm{Aug}_i(\boldsymbol{x}) | 1 \leq i \leq K\}$$

- Drawbacks
  - Optimizing the output divergence of each instance pair is inefficient, especially with large K
  - Some random augmentations which could cause significant semantic shift

- Conformal label distribution p

$$\sum_{k \in \mathcal{S}} p_k = 1 \text{ and } p_k = 0, \forall k \notin \mathcal{S}$$

$$\boldsymbol{\Psi}(\boldsymbol{x}, \mathcal{S}) = \sum_{\boldsymbol{z} \in \mathcal{A}(\boldsymbol{x})} \mathrm{KL}(\boldsymbol{p} || g(\boldsymbol{z})). \qquad (4)$$

Obtain the conformal label distribution

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

- Given this regularization term, we need to obtain the conformal label distribution of each instance before optimization the network's parameters
- p could be treated as latent variable and optimized simultaneously with the parameters θ from the objective function.
- Bi-level optimization.

$$\begin{aligned} &\arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{p}^*) \\ &\text{subject to } \boldsymbol{p}^* = \arg\min_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{p}); \\ &\qquad \textstyle\sum_{k\in\mathcal{S}} p_k = 1;\ p_k = 0, \forall k \notin \mathcal{S}. \end{aligned} \tag{5}$$

$$p_k^* = \frac{\left(\Pi_{\boldsymbol{z}\in\mathcal{A}(\boldsymbol{x})} g_k(\boldsymbol{z})\right)^{\frac{1}{|\mathcal{A}(\boldsymbol{x})|}}}{\sum_{j\in\mathcal{S}} \left(\Pi_{\boldsymbol{z}\in\mathcal{A}(\boldsymbol{x})} g_j(\boldsymbol{z})\right)^{\frac{1}{|\mathcal{A}(\boldsymbol{x})|}}}, \tag{7}$$

Method

- Loss

$$\mathcal{L}(\boldsymbol{x}, \mathcal{S}) = \mathcal{L}_{\text{Sup}}(\boldsymbol{x}, \mathcal{S}) + \gamma(t) \cdot \boldsymbol{\Psi}(\boldsymbol{x}, \mathcal{S}), \qquad (8)$$

- Dynamic balancing

$$\gamma(t) = \min\{\frac{t}{T'}\lambda, \lambda\}$$

Algorithm

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Algorithm 1 Our Regularized Training Method

Input: Training dataset $\mathcal{D} = \{x_i, S_i\}_{i=1}^{n}$;
The classifier $\boldsymbol{g}$ and its initial parameters $\boldsymbol{\theta}$;
Epochs T and iterations I;
The number of augmentations K;
Maximum balancing factor λ;

Procedure:

Initialize $\boldsymbol{p}$ for each instance by Eq.(6).
for $t = 1$ **to** T **do**
  for $i = 1$ **to** I **do**
    Fetch a random batch $\mathcal{B}$ from $\mathcal{D}$;
    Obtain the conformal label distributions by Eq.(7);
    Calculate the loss on the current batch by Eq. (8);
    Update network parameter $\boldsymbol{\theta}$ via gradient descent;
  end for
end for

Output: Learned multi-class classifier $\boldsymbol{g}$.

Experiments

Dataset	Method	$q = 0.1$	$q = 0.3$	$q = 0.5$	$q = 0.7$
Fashion-MNIST	Ours	$\mathbf{93.79 \pm 0.05\%}$	$\mathbf{93.72 \pm 0.20\%}$	$\mathbf{93.38 \pm 0.08\%}$	$\mathbf{92.19 \pm 0.03\%}$
	PiCO	$93.36 \pm 0.09\%$	$93.41 \pm 0.08\%$	$92.88 \pm 0.03\%$	$91.73 \pm 0.07\%$
	PRODEN	$89.15 \pm 0.58\%$	$89.10 \pm 0.26\%$	$88.22 \pm 0.35\%$	$85.87 \pm 0.28\%$
	VALEN	$86.52 \pm 0.16\%$	$85.80 \pm 0.17\%$	$85.01 \pm 0.30\%$	$81.92 \pm 0.30\%$
	LWS	$91.44 \pm 0.13\%$	$91.85 \pm 0.14\%$	$90.59 \pm 0.18\%$	$89.46 \pm 0.16\%$
	RC	$92.64 \pm 0.14\%$	$92.08 \pm 0.03\%$	$92.01 \pm 0.04\%$	$90.83 \pm 0.35\%$
	CC	$92.26 \pm 0.12\%$	$91.75 \pm 0.04\%$	$90.92 \pm 0.06\%$	$89.73 \pm 0.23\%$
	Fully Supervised	$93.92 \pm 0.07\%$			
Kuzushiji-MNIST	Ours	$\mathbf{98.27 \pm 0.07\%}$	$\mathbf{98.08 \pm 0.03\%}$	$\mathbf{97.44 \pm 0.04\%}$	$\mathbf{95.93 \pm 0.11\%}$
	PiCO	$97.68 \pm 0.06\%$	$97.34 \pm 0.07\%$	$97.15 \pm 0.03\%$	$91.90 \pm 0.04\%$
	PRODEN	$94.61 \pm 0.39\%$	$93.08 \pm 0.46\%$	$90.15 \pm 0.51\%$	$81.10 \pm 0.78\%$
	VALEN	$85.59 \pm 0.42\%$	$82.96 \pm 0.26\%$	$78.13 \pm 0.89\%$	$68.09 \pm 0.80\%$
	LWS	$96.22 \pm 0.10\%$	$96.15 \pm 0.24\%$	$95.43 \pm 0.02\%$	$93.63 \pm 0.04\%$
	RC	$96.84 \pm 0.09\%$	$96.31 \pm 0.15\%$	$96.17 \pm 0.06\%$	$95.84 \pm 0.12\%$
	CC	$96.45 \pm 0.04\%$	$96.16 \pm 0.02\%$	$95.62 \pm 0.10\%$	$95.33 \pm 0.14\%$
	Fully Supervised	$98.31 \pm 0.05\%$			

Table 1. Accuracy (mean±std) comparisons on Fashion-MNIST, Kuzushiji-MNIST, SVHN and CIFAR-10 with uniform partial labels on different ambiguity levels. The best result among each column is highlighted in bold.

Experiments

SVHN	Ours	**97.56 ± 0.02%**	**97.19 ± 0.06%**	**97.58 ± 0.05%**	**95.46 ± 0.23%**
	PiCO	96.01 ± 0.02%	96.24 ± 0.09%	95.89 ± 0.06%	94.71 ± 0.17%
	PRODEN	96.35 ± 0.25%	96.07 ± 0.23%	95.61 ± 0.21%	94.35 ± 0.29%
	VALEN	90.80 ± 0.16%	90.92 ± 0.30%	89.19 ± 0.64%	32.72 ± 0.17%
	LWS	96.12 ± 0.05%	95.72 ± 0.18%	33.75 ± 0.09%	19.59 ± 0.11%
	RC	96.20 ± 0.03%	95.54 ± 0.03%	96.03 ± 0.05%	95.73 ± 0.08%
	CC	96.15 ± 0.02%	95.79 ± 0.05%	95.35 ± 0.03%	94.55 ± 0.15%
	Fully Supervised	97.58 ± 0.03%			
CIFAR-10	Ours	**97.45 ± 0.04%**	**97.28 ± 0.02%**	**97.05 ± 0.05%**	**95.77 ± 0.08%**
	PiCO	95.78 ± 0.05%	95.25 ± 0.06%	94.73 ± 0.11%	92.73 ± 0.08%
	VALEN	68.97 ± 0.23%	63.61 ± 0.17%	52.57 ± 0.32%	38.02 ± 2.16%
	PRODEN	91.94 ± 0.32%	91.10 ± 0.50%	89.82 ± 0.47%	86.48 ± 0.47%
	LWS	86.47 ± 0.20%	84.31 ± 0.14%	54.73 ± 0.19%	38.49 ± 0.24%
	RC	88.96 ± 0.06%	87.49 ± 0.17%	83.48 ± 0.19%	75.01 ± 0.19%
	CC	88.78 ± 0.05%	86.69 ± 0.40%	83.75 ± 0.28%	77.60 ± 0.22%
	Fully Supervised	97.57 ± 0.03%			

Table 1. Accuracy (mean±std) comparisons on Fashion-MNIST, Kuzushiji-MNIST, SVHN and CIFAR-10 with uniform partial labels on different ambiguity levels. The best result among each column is highlighted in bold.

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Table 2. Accuracy (mean±std) comparisons on CIFAR-100 with uniform partial labels on different ambiguity levels.

Dataset	Method	$q = 0.01$	$q = 0.05$	$q = 0.1$	$q = 0.2$
CIFAR-100	Ours	$\mathbf{83.12 \pm 0.20\%}$	$\mathbf{82.77 \pm 0.10\%}$	$\mathbf{82.24 \pm 0.07\%}$	$\mathbf{80.97 \pm 0.29\%}$
	PiCO	$74.39 \pm 0.15\%$	$73.97 \pm 0.09\%$	$51.94 \pm 0.11\%$	$20.29 \pm 0.04\%$
	PRODEN	$72.55 \pm 0.77\%$	$71.55 \pm 0.94\%$	$70.84 \pm 0.87\%$	$58.86 \pm 0.85\%$
	VALEN	$43.68 \pm 0.50\%$	$43.19 \pm 0.56\%$	$35.00 \pm 0.57\%$	$4.51 \pm 0.61\%$
	LWS	$58.54 \pm 0.12\%$	$55.19 \pm 0.23\%$	$40.12 \pm 0.34\%$	$23.90 \pm 0.18\%$
	RC	$64.95 \pm 0.23\%$	$62.48 \pm 0.14\%$	$57.48 \pm 0.04\%$	$44.13 \pm 0.23\%$
	CC	$63.74 \pm 0.17\%$	$61.22 \pm 0.21\%$	$58.65 \pm 0.06\%$	$51.65 \pm 0.49\%$
	Fully Supervised	$83.16 \pm 0.19\%$			

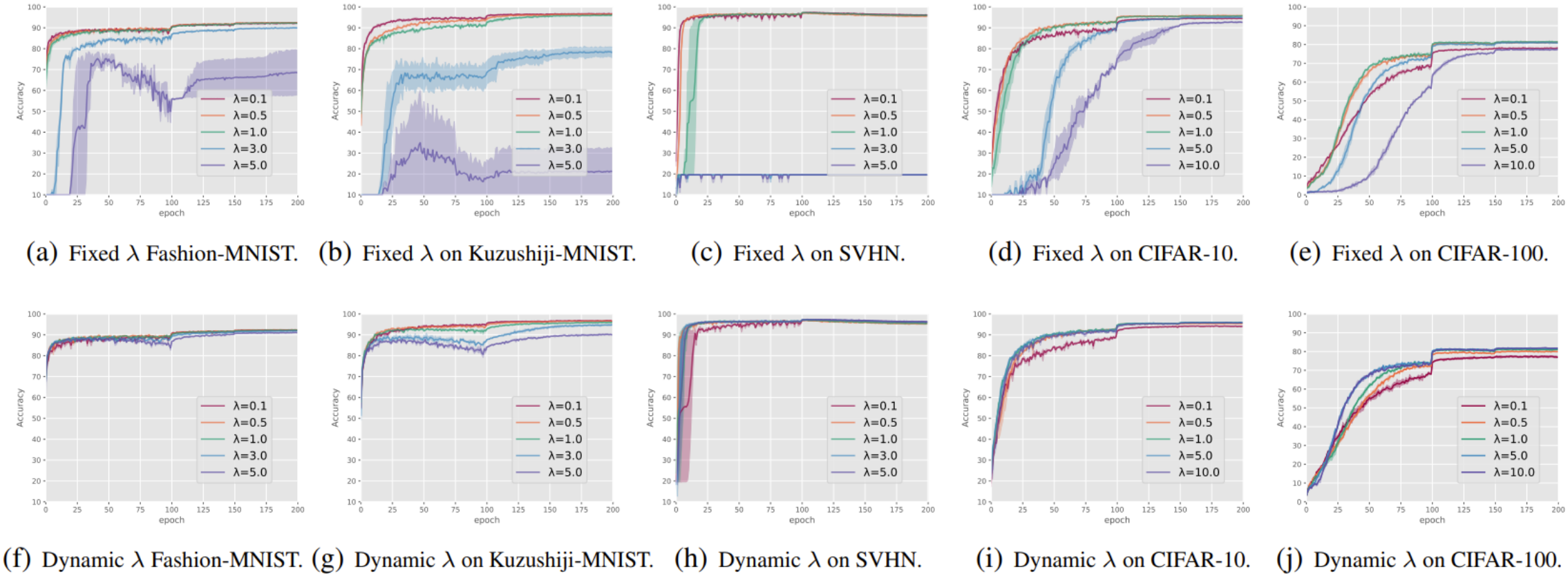
Experiments

- Instance-dependent partial labels
- The flipping probability of each incorrect label is computed by $q_j(x) = \frac{\hat{g}_j(x)}{max_{k \in S} \hat{g}_k(x)}$
- $\hat{g}$ denotes a pre-trained neural network.

Table 3. Accuracy (mean±std) comparisons on Kuzushiji-MNIST, Fashion-MNIST, CIFAR-10 with instance-dependent partial labels.

Method	Kuzushiji-MNIST	Fashion-MNIST	CIFAR10
Ours	**95.07 ± 0.06%**	**89.21 ± 0.21%**	87.80 ± 0.11%
PiCO	92.87 ± 0.08%	86.93 ± 0.20%	**92.64 ± 0.07%**
PRODEN	87.71 ± 0.62%	84.25 ± 0.61%	76.51 ± 0.69%
VALEN	83.16 ± 0.33%	85.40 ± 0.10%	62.37 ± 0.35%
LWS	91.17 ± 0.18%	86.14 ± 0.34%	44.08 ± 0.15%
RC	94.00 ± 0.12%	89.10 ± 0.24%	75.90 ± 0.34%
CC	94.01 ± 0.05%	88.33 ± 0.17%	79.58 ± 0.22%

Experiments(Ablation Study)



(a) Fixed λ Fashion-MNIST. (b) Fixed λ on Kuzushiji-MNIST. (c) Fixed λ on SVHN. (d) Fixed λ on CIFAR-10. (e) Fixed λ on CIFAR-100.

(f) Dynamic λ Fashion-MNIST. (g) Dynamic λ on Kuzushiji-MNIST. (h) Dynamic λ on SVHN. (i) Dynamic λ on CIFAR-10. (j) Dynamic λ on CIFAR-100.

Figure 2. Accuracy curves with different balancing strategies. The top row shows results with fixed λ and the bottom row shows the results with dynamic λ. Dark colors show the mean accuracy of 5 trials and light colors show standard deviation.

Experiments(Ablation Study)

- The importance of regularization term.

Table 4. Accuracy (mean±std) of the degenerated method with uniform partial labels.

	$q = 0.5$	$q = 0.7$
Fashion-MNIST	$91.49 \pm 0.71\%$ $\downarrow$ **(1.89%)**	$90.61 \pm 0.66\%$ $\downarrow$ **(1.58%)**
Kuzushiji-MNIST	$95.99 \pm 0.06\%$ $\downarrow$ **(1.45%)**	$92.28 \pm 0.20\%$ $\downarrow$ **(3.65%)**
SVHN	$96.60 \pm 0.71\%$ $\downarrow$ **(0.98%)**	$95.02 \pm 0.09\%$ $\downarrow$ **(0.44%)**
CIFAR-10	$95.39 \pm 1.47\%$ $\downarrow$ **(1.66%)**	$93.29 \pm 1.23\%$ $\downarrow$ **(2.48%)**
	$q = 0.05$	$q = 0.1$
CIFAR-100	$76.07 \pm 0.15\%$ $\downarrow$ **(6.70%)**	$74.93 \pm 0.19\%$ $\downarrow$ **(7.31%)**

$$p_k^* = \frac{\left(\prod_{z \in \mathcal{A}(x)} g_k(z)\right)^{\frac{1}{|\mathcal{A}(x)|}}}{\sum_{j \in \mathcal{S}} \left(\prod_{z \in \mathcal{A}(x)} g_j(z)\right)^{\frac{1}{|\mathcal{A}(x)|}}}, \tag{7}$$

- Specifically, we replace the conformal label distribution inferred by Eq. (7) with simply re-normalized model outputs, and remaining the augmentation techniques.

Experiments(Ablation Study)

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

- The influence of backbone network.

Table 5. Accuracy (mean±std) comparison between PiCO and Our method with different backbones. WRN-34-10 is short for Wide-ResNet-34-10.

		PiCO	Ours
ResNet-18	Fully Supervised	$73.56 \pm 0.10\%$	$79.88 \pm 0.03\%$
	$p = 0.01$	$73.09 \pm 0.34\%$ $\downarrow$ **(0.47%)**	$79.54 \pm 0.12\%$ $\downarrow$ **(0.34%)**
	$p = 0.05$	$72.74 \pm 0.30\%$ $\downarrow$ **(0.82%)**	$78.96 \pm 0.06\%$ $\downarrow$ **(0.92%)**
	$p = 0.10$	$69.91 \pm 0.24\%$ $\downarrow$ **(3.65%)**	$77.72 \pm 0.08\%$ $\downarrow$ **(2.15%)**
WRN-34-10	Fully Supervised	$74.47 \pm 0.14\%$	$83.16 \pm 0.19\%$
	$p = 0.01$	$74.39 \pm 0.15\%$ $\downarrow$ **(0.08%)**	$83.12 \pm 0.20\%$ $\downarrow$ **(0.04%)**
	$p = 0.05$	$73.97 \pm 0.09\%$ $\downarrow$ **(0.50%)**	$82.77 \pm 0.10\%$ $\downarrow$ **(0.39%)**
	$p = 0.10$	$51.94 \pm 0.11\%$ $\downarrow$ **(22.53%)**	$82.24 \pm 0.07\%$ $\downarrow$ **(0.92%)**

Thanks