Towards Realistic Long-Tailed Semi-Supervised Learning: Consistency Is All You Need

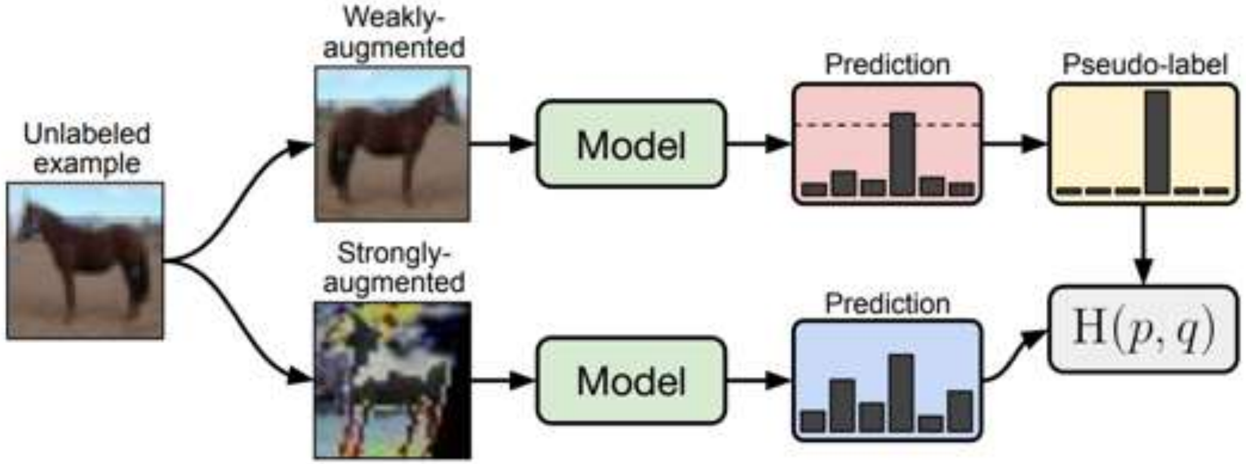
Tong Wei, Kai Gan
School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
{weit, gank}@seu.edu.cn

TUE-AM-330

Long-Tailed Semi-Supervised Learning (LTSSL)

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Common semi-supervised learning method: FixMatch



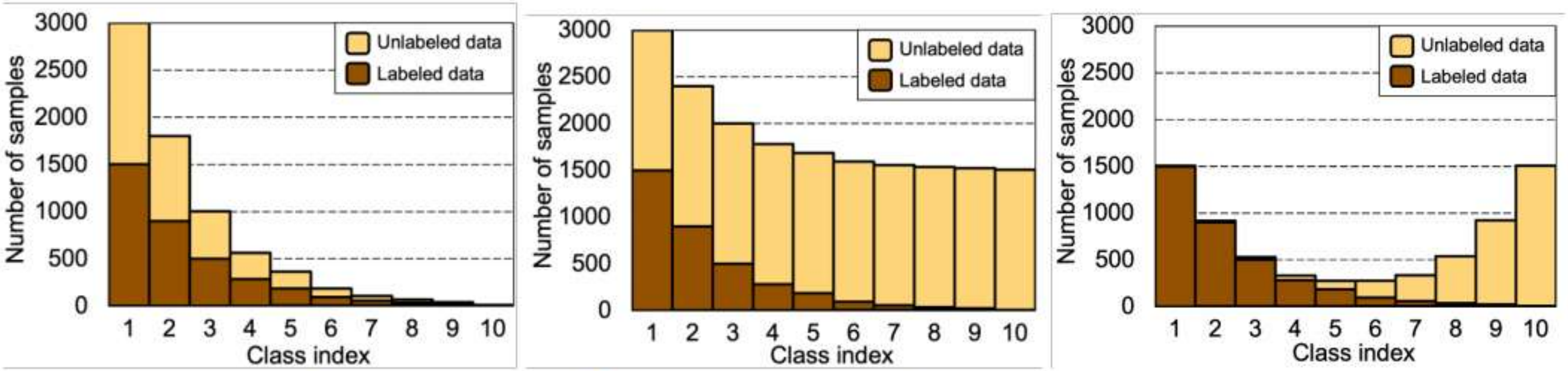
Labeled(Cross-entropy)
+
Unlabeled(Consistency regularization)

$$\min_{\theta\in\Theta} \underbrace{\sum_{i=1}^{N} \ell(f(x_i^{(l)};\theta), y_i^{(l)})}_{\text{supervised } (\mathcal{L}_{\text{labeled}})} + \underbrace{\sum_{j=1}^{M} \Omega(x_j^{(u)};\theta)}_{\text{unsupervised}},$$

Recent progress on SSL has revealed promising performance in various tasks

However, most existing SSL algorithms assume the datasets are class-balanced

Three typical types of class distribution of unlabeled data

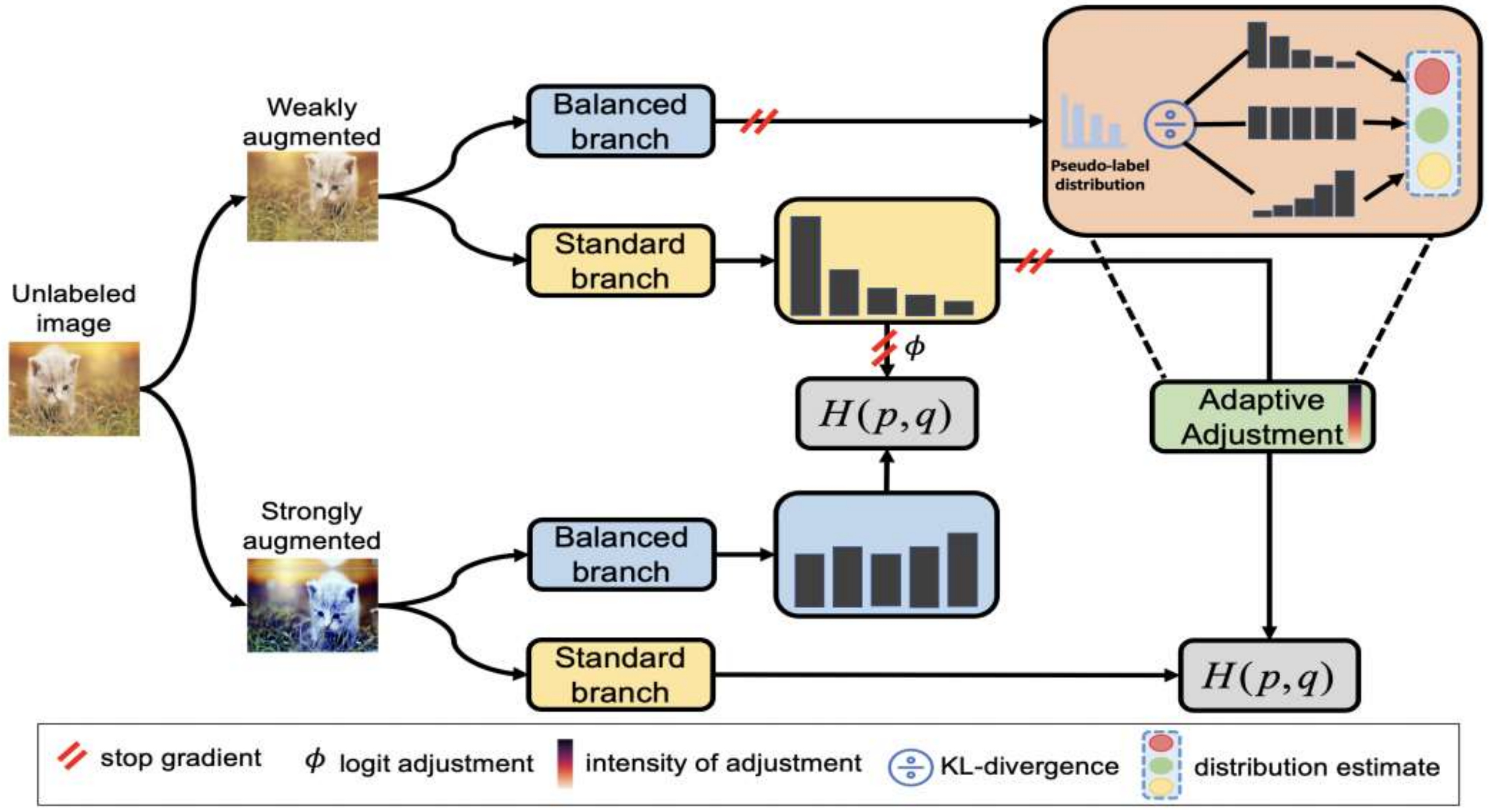


(a) *Consistent* class distribution (b) *Uniform* class distribution (c) *Reversed* class distribution

Adaptive Consistency Regularizer (ACR)

Two findings:

1) Pseudo-labels biased towards minority classes can benefit the classifier learning;
2) Pseudo-label distribution that approximates the true distribution helps learn better feature extractor.



Balanced branch

Adjust pseudo-labels appropriately biased toward the minority class via logit adjustment

Standard branch

Refine the original pseudo-labels to match the true class distribution of unlabeled data and enhance their accuracy

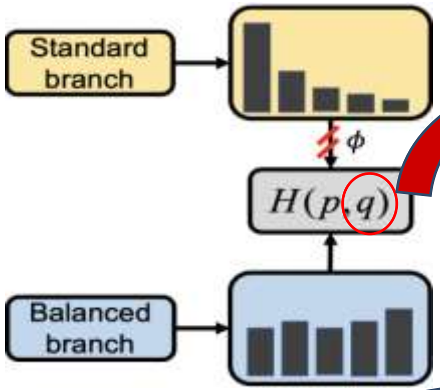
Adaptive Consistency Regularizer (ACR)

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

For labeled data

- Standard branch: Cross-entropy
- Balanced branch: Balanced Softmax

$$\ell(y, f(x)) = -\log \frac{e^{f_y(x)+\tau\cdot\log\pi_y}}{\sum_{y'\in[L]} e^{f_{y'}(x)+\tau\cdot\log\pi_{y'}}}$$

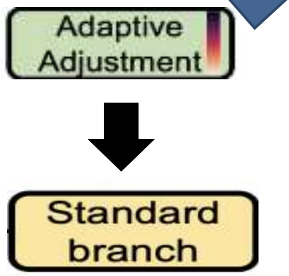


$$\tilde{q}(x_j^{(u)}) = \arg\max f\left(x_j^{(u)}\right) - \tau\cdot\log\pi.$$

For unlabeled data

- Balanced branch
- Standard branch

Estimated distribution for unlabeled data by EMA



Consistency Loss

Adaptive Consistency Regularizer (ACR)

π_{con} Anchor distribution for consistent

π_{uni} Anchor distribution for uniform

π_{rev} Anchor distribution for reversed

π_{est} Estimated distribution for unlabeled data

Calculate distances to each anchor distribution

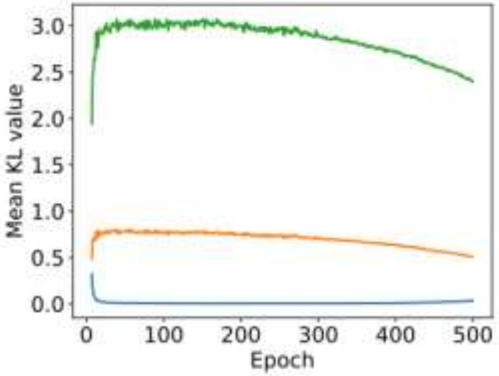
$$dist_{\text{con}} = \frac{1}{2}\left(D_{KL}\left(\pi_{\text{con}} \| \pi_{\text{est}}\right) + D_{KL}\left(\pi_{\text{est}} \| \pi_{\text{con}}\right)\right)$$

$$dist_{\text{uni}} = \frac{1}{2}\left(D_{KL}\left(\pi_{\text{uni}} \| \pi_{\text{est}}\right) + D_{KL}\left(\pi_{\text{est}} \| \pi_{\text{uni}}\right)\right)$$

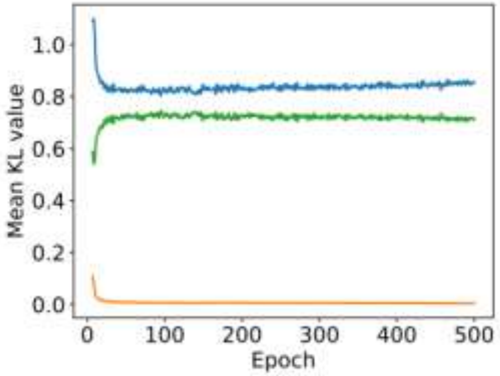
$$dist_{\text{rev}} = \frac{1}{2}\left(D_{KL}\left(\pi_{\text{rev}} \| \pi_{\text{est}}\right) + D_{KL}\left(\pi_{\text{est}} \| \pi_{\text{rev}}\right)\right),$$

$$\tau(t) = \frac{2e^{dist_{\text{con}}^{(t-1)}}}{e^{dist_{\text{con}}^{(t-1)}} + e^{dist_{\text{uni}}^{(t-1)}} + e^{dist_{\text{rev}}^{(t-1)}}}$$

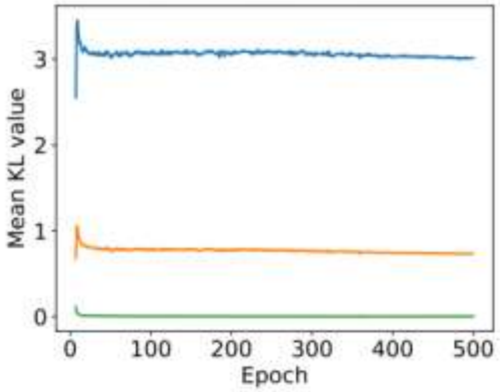
Consistent — Uniform — Reversed



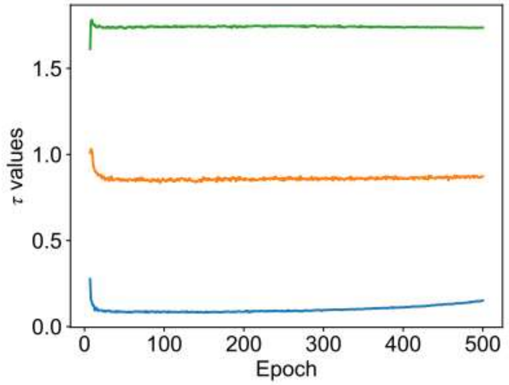
(a) Distance for *consistent* setting



(b) Distance for *uniform* setting



(c) Distance for *reversed* setting



(d) τ values for LA

Adaptive Consistency Regularizer (ACR)

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Sample mask generation

$$\mathcal{L}_{\text{b-con}} = \sum_{j=1}^{M} \tilde{M}(x_j^{(u)})\ell\left(\tilde{f}(\mathcal{A}(x_j^{(u)})), \tilde{q}_{j_\delta}\right),$$

$$\tilde{M}(x_j^{(u)}) = \mathbb{I}\left(\max\left(\delta(\tilde{f}(x_j^{(u)}))\right) \geq \rho\right) \quad \vee$$
$$\mathbb{I}\left(\max\left(\delta(f(x_j^{(u)}) - \tau \cdot \log\pi)\right) \geq \rho\right),$$

$\tilde{M}(x_j^{(u)})$	Sample mask for $x_j^{(u)}$ in balanced branch
δ	Softmax function
$\mathbb{I}$	Indicator function
ρ	Predefined threshold
π	Distribution of labeled data

In this way, we can

(1) select more samples for the minority classes by considering the balanced branch's output;

(2) obtain more confident samples through the newly constructed sample mask, which is beneficial for consistency loss to work.

Experimental Results

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Test accuracy for consistent setting

Algorithm	CIFAR10-LT				CIFAR100-LT			
	$\gamma=\gamma_l=\gamma_u=100$		$\gamma=\gamma_l=\gamma_u=150$		$\gamma=\gamma_l=\gamma_u=10$		$\gamma=\gamma_l=\gamma_u=20$	
	$N_1=500$ $M_1=4000$	$N_1=1500$ $M_1=3000$	$N_1=500$ $M_1=4000$	$N_1=1500$ $M_1=3000$	$N_1=50$ $M_1=400$	$N_1=150$ $M_1=300$	$N_1=50$ $M_1=400$	$N_1=150$ $M_1=300$
Supervised	47.3 ±0.95	61.9 ±0.41	44.2 ±0.33	58.2 ±0.29	29.6 ±0.57	46.9 ±0.22	25.1 ±1.14	41.2 ±0.15
w/ LA [22]	53.3 ±0.44	53.3 ±0.21	49.5 ±0.40	67.1 ±0.78	30.2 ±0.44	48.7 ±0.89	26.5 ±1.31	44.1 ±0.42
FixMatch [29]	67.8 ±1.13	77.5 ±1.32	62.9 ±0.36	72.4 ±1.03	45.2 ±0.55	56.5 ±0.06	40.0 ±0.96	50.7 ±0.25
w/ DARP [14]	74.5 ±0.78	77.8 ±0.63	67.2 ±0.32	73.6 ±0.73	49.4 ±0.20	58.1 ±0.44	43.4 ±0.87	52.2 ±0.66
w/ CReST+ [34]	76.3 ±0.86	78.1 ±0.42	67.5 ±0.45	73.7 ±0.34	44.5 ±0.94	57.4 ±0.18	40.1 ±1.28	52.1 ±0.21
w/ DASO [25]	76.0 ±0.37	79.1 ±0.75	70.1 ±1.81	75.1 ±0.77	49.8 ±0.24	59.2 ±0.35	43.6 ±0.09	52.9 ±0.42
FixMatch+LA [22]	75.3 ±2.45	82.0 ±0.36	67.0 ±2.49	78.0 ±0.91	47.3 ±0.42	58.6 ±0.36	41.4 ±0.93	53.4 ±0.32
w/ DARP [14]	76.6 ±0.92	80.8 ±0.62	68.2 ±0.94	76.7 ±1.13	50.5 ±0.78	59.9 ±0.32	44.4 ±0.65	53.8 ±0.43
w/ CReST+ [34]	76.7 ±1.13	81.1 ±0.57	70.9 ±1.18	77.9 ±0.71	44.0 ±0.21	57.1 ±0.55	40.6 ±0.55	52.3 ±0.20
w/ DASO [25]	77.9 ±0.88	82.5 ±0.08	70.1 ±1.68	79.0 ±2.23	50.7 ±0.51	60.6 ±0.71	44.1 ±0.61	55.1 ±0.72
FixMatch+ABC [18]	78.9 ±0.82	83.8 ±0.36	66.5 ±0.78	80.1 ±0.45	47.5 ±0.18	59.1 ±0.21	41.6 ±0.83	53.7 ±0.55
w/ DASO [25]	80.1 ±1.16	83.4 ±0.31	70.6 ±0.80	80.4 ±0.56	50.2 ±0.62	60.0 ±0.32	44.5 ±0.25	55.3 ±0.53
FixMatch w/ ACR (ours)	**81.6** ±0.19	**84.1** ±0.39	**77.0** ±1.19	**80.9** ±0.22	**55.7** ±0.12	**65.6** ±0.16	**48.0** ±0.75	**58.9** ±0.36

ACR outperforms all algorithms even though most these methods are particularly developed based on the assumption that labeled and unlabeled data share the same class distribution

Experimental Results

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Test accuracy for inconsistent settings

Algorithm	CIFAR10-LT ($\gamma_l \neq \gamma_u$)				STL10-LT (γ_u = N/A)			
	$\gamma_u = 1$ (uniform)		$\gamma_u = 1/100$ (reversed)		$\gamma_l = 10$		$\gamma_l = 20$	
	$N_1 = 500$ $M_1 = 4000$	$N_1 = 1500$ $M_1 = 3000$	$N_1 = 500$ $M_C = 4000$	$N_1 = 1500$ $M_C = 3000$	$N_1 = 150$ $M = 100k$	$N_1 = 450$ $M = 100k$	$N_1 = 150$ $M = 100k$	$N_1 = 450$ $M = 100k$
FixMatch [29]	73.0 ±3.81	81.5 ±1.15	62.5 ±0.94	71.8 ±1.70	56.1 ±2.32	72.4 ±0.71	47.6 ±4.87	64.0 ±2.27
w/ DARP [14]	82.5 ±0.75	84.6 ±0.34	70.1 ±0.22	80.0 ±0.93	66.9 ±1.66	75.6 ±0.45	59.9 ±2.17	72.3 ±0.60
w/ CReST [34]	83.2 ±1.67	87.1 ±0.28	70.7 ±2.02	80.8 ±0.39	61.7 ±2.51	71.6 ±1.17	57.1 ±3.67	68.6 ±0.88
w/ CReST+ [34]	82.2 ±1.53	86.4 ±0.42	62.9 ±1.39	72.9 ±2.00	61.2 ±1.27	71.5 ±0.96	56.0 ±3.19	68.5 ±1.88
w/ DASO [25]	86.6 ±0.84	88.8 ±0.59	71.0 ±0.95	80.3 ±0.65	70.0 ±1.19	78.4 ±0.80	65.7 ±1.78	75.3 ±0.44
w/ ACR (ours)	**92.1** ±0.18	**93.5** ±0.11	**85.0** ±0.09	**89.5** ±0.17	**77.1** ±0.24	**83.0** ±0.32	**75.1** ±0.70	**81.5** ±0.25

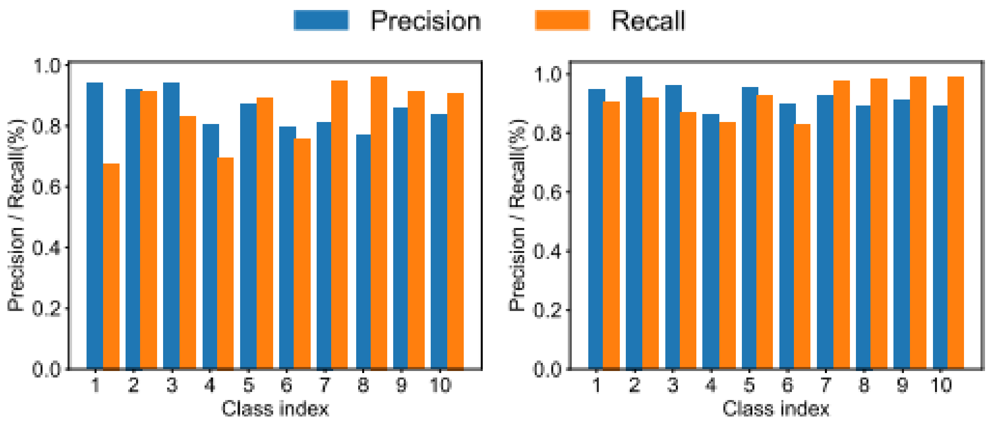
Algorithm	CIFAR100-LT ($\gamma_l \neq \gamma_u$)			
	$\gamma_u = 1$ (uniform)		$\gamma_u = 1/10$ (reversed)	
	$N_1 = 50$ $M_1 = 400$	$N_1 = 150$ $M_1 = 300$	$N_1 = 50$ $M_C = 400$	$N_1 = 150$ $M_C = 300$
FixMatch [29]	45.5 ±0.71	58.1 ±0.72	44.2 ±0.43	57.3 ±0.19
w/ DARP [14]	43.5 ±0.95	55.9 ±0.32	36.9 ±0.48	51.8 ±0.92
w/ CReST [34]	43.5 ±0.30	59.2 ±0.25	39.0 ±1.11	56.4 ±0.62
w/ CReST+ [34]	43.6 ±1.60	58.7 ±0.16	39.1 ±0.77	56.4 ±0.78
w/ DASO [25]	53.9 ±0.66	61.8 ±0.98	51.0 ±0.19	60.0 ±0.31
w/ ACR (ours)	**66.0** ±0.25	**73.4** ±0.22	**57.0** ±0.46	**67.6** ±0.12

Test accuracy on ImageNet-127

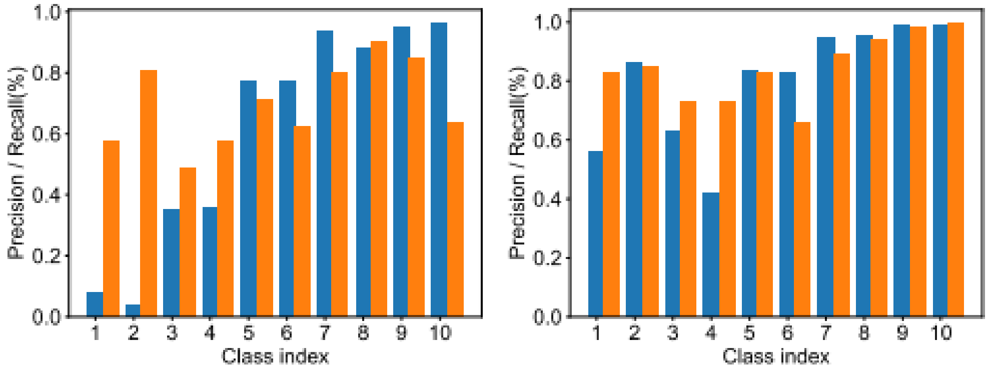
Algorithm	32×32	64×64
FixMatch [29]	29.7	42.3
w/ DARP [14]	30.5	42.5
w/ DARP+cRT [14]	39.7	51.0
w/ CReST+ [34]	32.5	44.7
w/ CReST++LA [22]	40.9	55.9
w/ CoSSL [9]	43.7	53.9
w/ TRAS [35]	46.2	54.1
w/ ACR (ours)	**57.2**	**63.6**

Experimental Results

The precision and recall of pseudo-labels

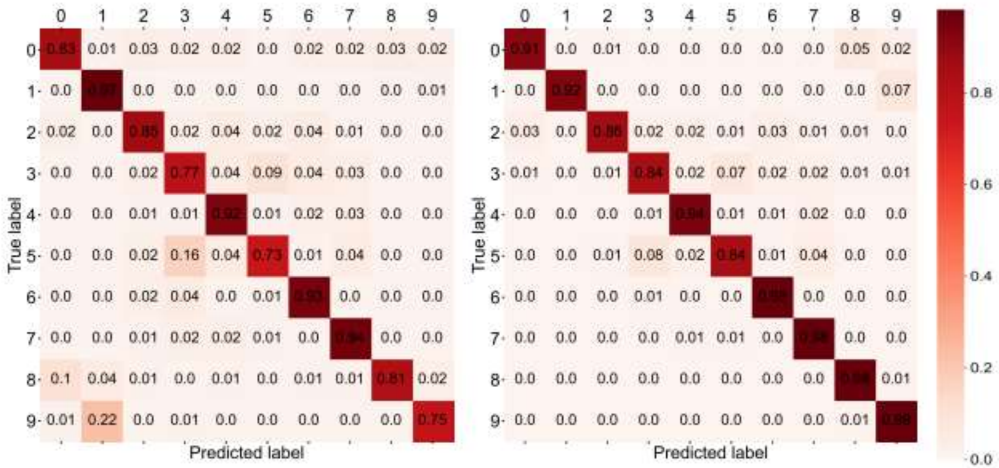


(a) DASO w/ LA for *uniform* (b) ACR for *uniform*

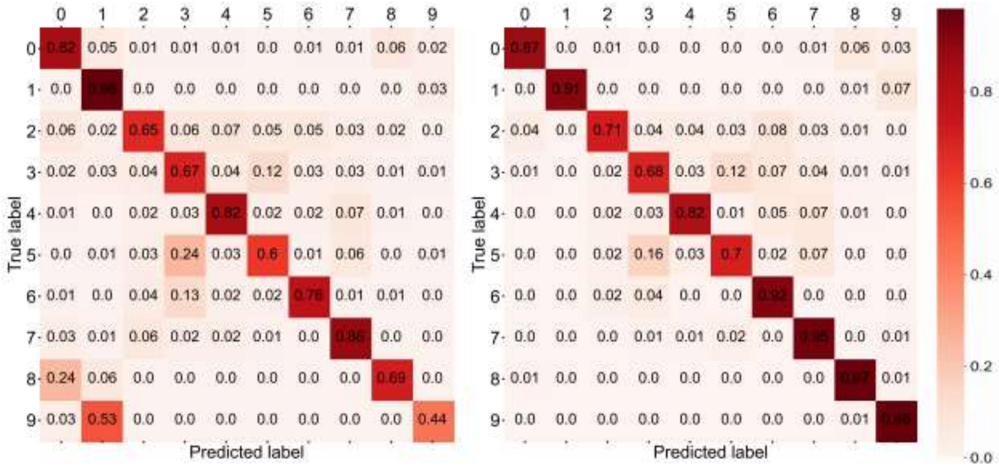


(c) DASO w/ LA for *reversed* (d) ACR for *reversed*

Confusion matrices



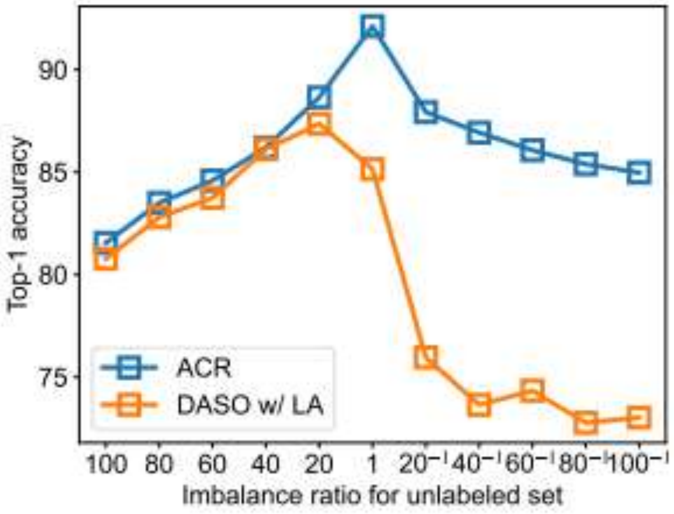
(a) DASO for *uniform* (b) ACR for *uniform*



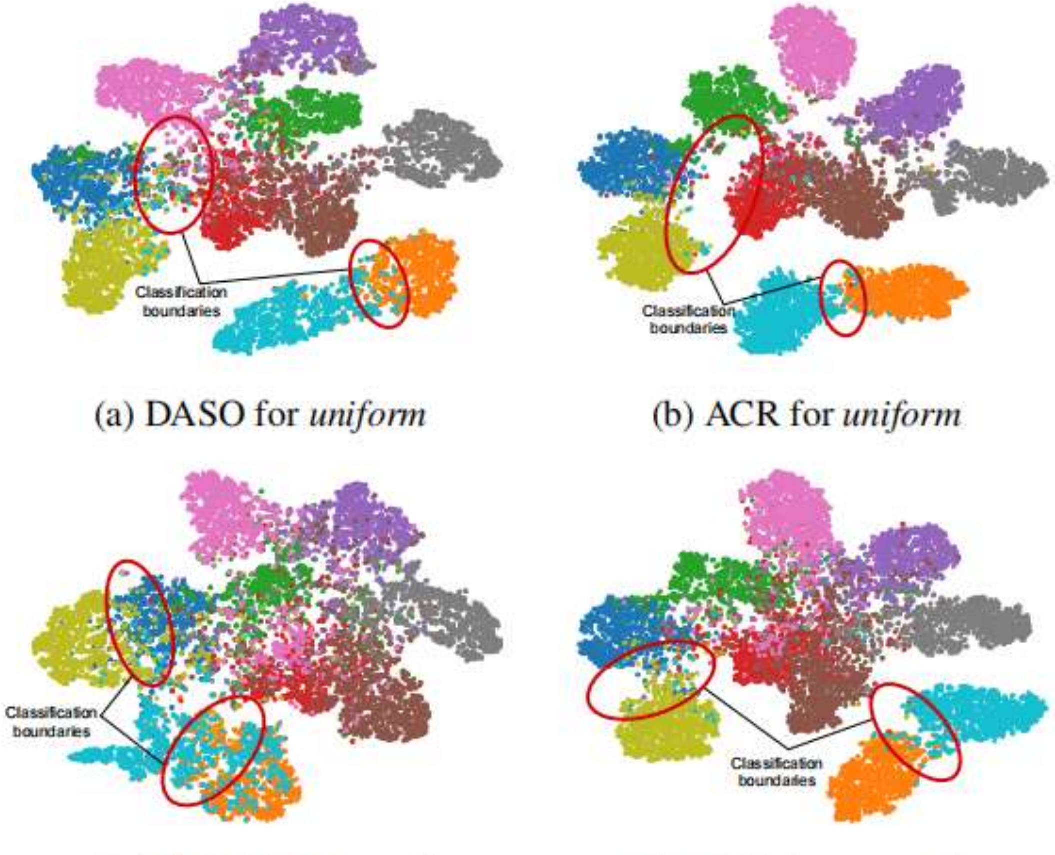
(c) DASO for *reversed* (d) ACR for *reversed*

Experimental Results

More settings:



The t-SNE visualization:



(a) DASO for *uniform* (b) ACR for *uniform*

(c) DASO for *reversed* (d) ACR for *reversed*

Ablation studies:

Ablations	CIFAR10-LT			CIFAR100-LT		
	Con	Uni	Rev	Con	Uni	Rev
ACR(ours)	81.6	92.1	85.0	55.7	66.0	57.0
w/o sample mask principle	81.7	91.1	84.6	55.0	63.7	55.0
w/o adaptive LA	76.8	92.4	85.1	53.5	62.8	56.1
w/o LA for balanced branch	74.3	90.6	83.5	54.5	66.2	56.7
w/o balanced softmax	76.7	93.0	84.8	55.3	65.6	57.3
w/o gradients from balanced branch	73.7	92.3	85.2	54.3	65.2	56.7
w/o labeled data in unlabeled set	81.0	92.7	79.9	56.1	66.4	56.8

Conclusion

We presents a simple and effective method by minimizing the adaptive consistency regularizer (ACR) for long-tailed semi-supervised learning with unknown class distributions of the unlabeled data.

- Benefit classifier learning by generating pseudo-labels that are properly biased towards minority classes.
- Benefit representation learning by generating pseudo-labels whose distribution approximates the true class distribution.

Thanks!

gank@seu.edu.cn