南京航空航天大學

Nanjing University of Aeronautics and Astronautics

SimMatch: Semi-supervised Learning with Similarity Matching

Mingkai Zheng[1,2] Shan You[2*]
Lang Huang[3] Fei Wang[4] Chen Qian[2] Chang Xu[1]
[1]School of Computer Science, Faculty of Engineering, The University of Sydney
[2]SenseTime Research [3] The University of Tokyo
[4]University of Science and Technology of China

CVPR2022

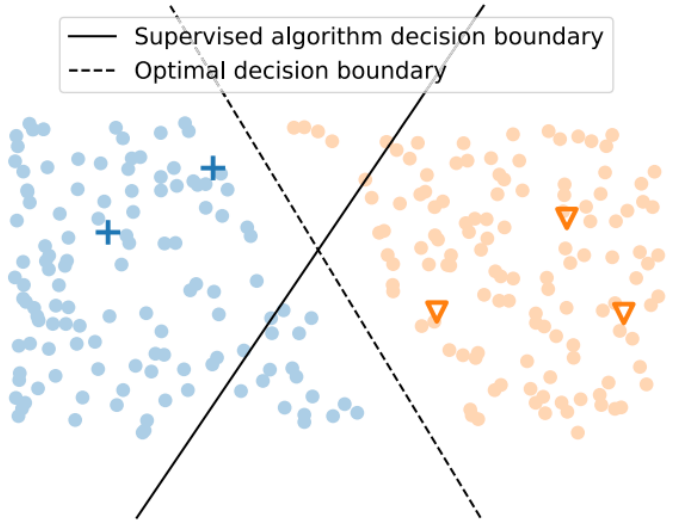
Background

Consistency Regularization(Smoothness)

The model's output should remain unchanged when the input is perturbed.

Entropy Minimization(Low-density)

The classifier's decision boundary should not pass through high-density regions of the marginal data distribution.



(a) Smoothness and low-density assumptions.

Self-supervised contrastive learning

Distinguish instances by their similarity.
Another form of **Consistency Regularization.**

$$-\log \frac{\exp(z(\mathrm{Aug}(x_i)) \cdot z(\mathrm{Aug}(x_i))/t)}{\sum_{j=1}^{N} \exp(z(\mathrm{Aug}(x_i)) \cdot z(\mathrm{Aug}(x_j))/t)}$$

How to use similarity of instances in semi-supervised learning task to increase its performance ?

Method

Labeled examples CE loss

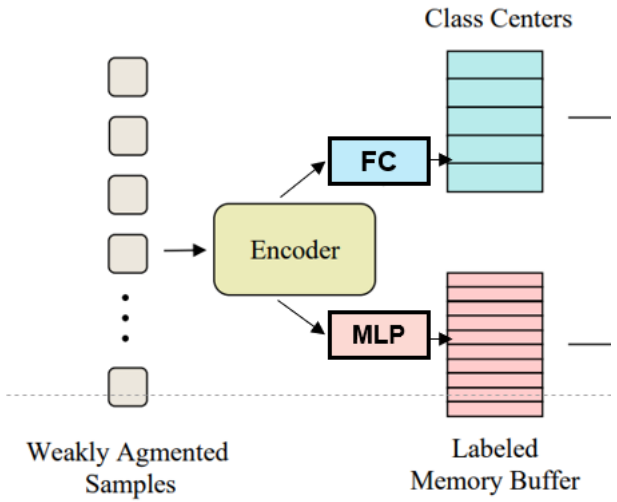
$$\mathcal{L}_s = \frac{1}{B}\sum \mathrm{H}(y, p)$$

Input weakly and strongly augmented **unlabeled** sample to the network to get prediction p^w and p^s, unsupervised classification loss can be defined as the cross entropy between these two predictions:

$$\mathcal{L}_u = \frac{1}{\mu B}\sum \mathbb{1}(\max DA(p^w) > \tau)\mathrm{H}(DA(p^w), p^s)$$

Like in FixMatch

$$DA(p^w) = Normalized(p^w / {p_{avg}}^w) \qquad \mathrm{Normalize}(x)_i = x_i / \sum_j x_j$$



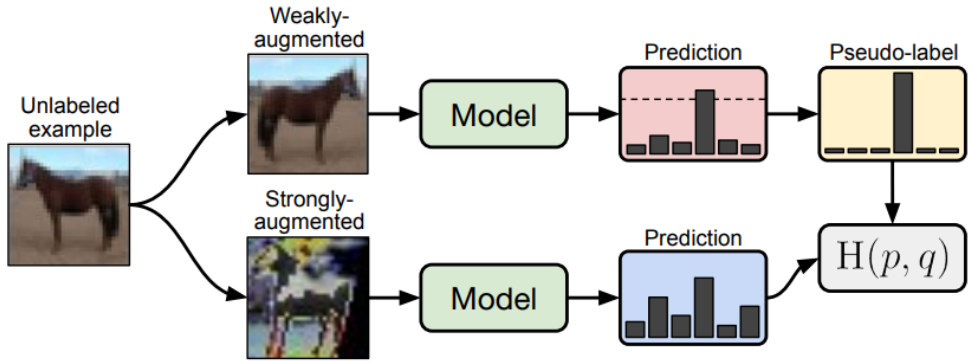


Figure 1: Diagram of FixMatch. A weakly-augmented image (top) is fed into the model to obtain predictions (red box). When the model assigns a probability to any class which is above a threshold (dotted line), the prediction is converted to a one-hot pseudo-label. Then, we compute the model's prediction for a strong augmentation of the same image (bottom). The model is trained to make its prediction on the strongly-augmented version match the pseudo-label via a cross-entropy loss.

Method

Distribution Alignment

Maximize the mutual information between the model's input and output for unlabeled data.

$$\mathcal{I}(y;x) = \iint p(y,x) \log \frac{p(y,x)}{p(y)p(x)} \, \mathrm{d}y \, \mathrm{d}x$$
$$= \mathcal{H}(\mathbb{E}_x[p_{\text{model}}(y|x;\theta)]) - \mathbb{E}_x[\mathcal{H}(p_{\text{model}}(y|x;\theta))]$$

Model should predict each class with equal frequency

Familiar entropy minimization objective

Method

Instance Similarity Matching

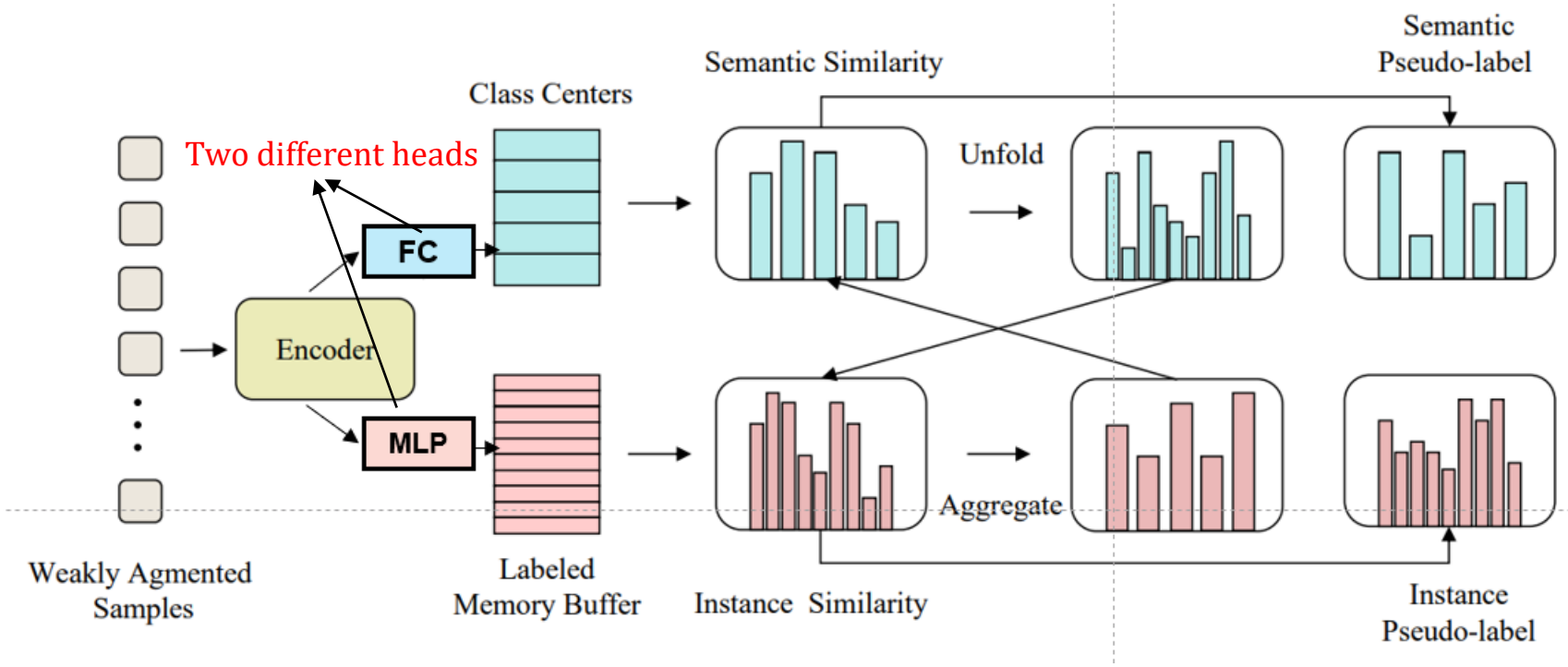
Use z_b^w and z_b^s to denote **unlabeled** sample embedding from weakly and strongly augmented view,

Labeled Memory Buffer stores K weakly augmented embeddings of labeled samples.

Use softmax layer to calculate similarities：

$$q_i^w = \frac{\exp(sim(\mathbf{z}_b^w, \mathbf{z}_i)/t)}{\sum_{k=1}^{K} \exp(sim(\mathbf{z}_b^w, \mathbf{z}_k)/t)} \qquad q_i^s = \frac{\exp(sim(\mathbf{z}_b^s, \mathbf{z}_i)/t)}{\sum_{k=1}^{K} \exp(sim(\mathbf{z}_b^s, \mathbf{z}_k)/t)}$$

t is the temperature parameter that controls the sharpness of the distribution. (From MixMatch)



Consistency regularization loss (Self-supervised contrastive learning)

$$\mathcal{L}_{in} = \frac{1}{\mu B} \sum \mathrm{H}(q^w, q^s)$$

Total loss is:

$$\mathcal{L}_{overall} = \mathcal{L}_s + \lambda_u \mathcal{L}_u + \lambda_{in} \mathcal{L}_{in}$$

Method

Label Propagation through SimMatch

The generation of the **instance** pseudo-labels q^w are still in a fully unsupervised manner.

How to use labeled information sufficiently?

Given a weakly augmented unlabeled sample, compute the semantic similarity $p^w \in \mathbf{R}^{1\times L}$ and instance similarity $q^w \in \mathbf{R}^{1\times K}$

L is the number of classes , and K is the number of labeled samples .

semantic similarity space (L dimension) ⇄ instance similarity space (K dimension)

Unfold

Aggregate

Unfold

$$p_j^{unfold} = p_i^w, where\ \text{class}\left(q_j^w\right) = \text{class}\left(p_i^w\right)$$

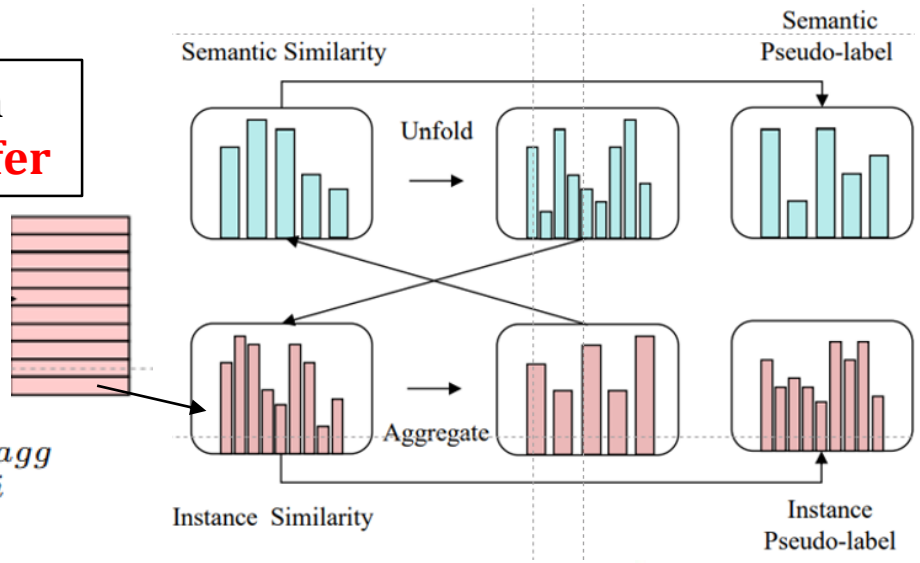
J iterates from 0 to K in **Labeled Memory Buffer**

$$\widehat{q}_i = \frac{q_i^w p_i^{unfold}}{\sum_{k=1}^{K} q_k^w p_k^{unfold}}$$

Replace the old one q^w to generate **Instance Pseudo–label**

Aggregate

$$q_i^{agg} = \sum_{j=0}^{K} \mathbb{1}(class(p_i^w) = class(q_j^w)) q_j^w \qquad \widehat{p}_i = \alpha p_i^w + (1-\alpha) q_i^{agg}$$



Method

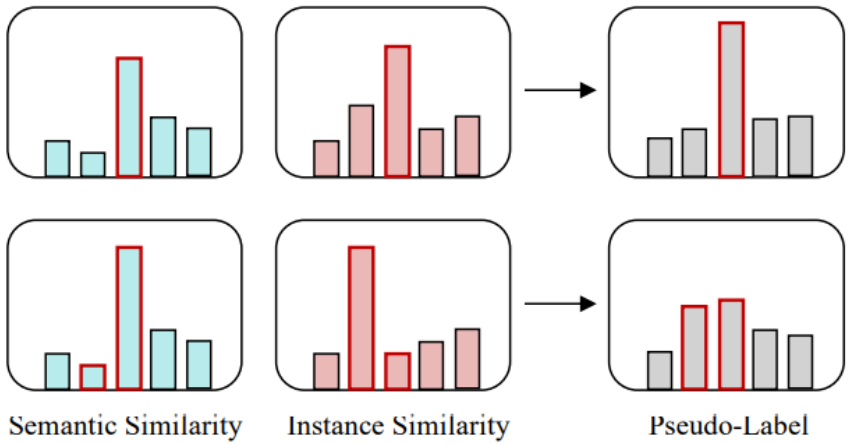


Figure 3. The intuition behind label propagation. If the semantic and instance similarities are similar, the result pseudo-label will be much sharper and produce high confidence for some classes. When these two similarities are different, the result pseudo-label will be much flatter.

Feature Memory Buffer

$Q_f \in \mathbf{R}^{K\times D}$

64M GPU space

Label Memory Buffer

$Q_l \in \mathbf{R}^{K\times 1}$

1M GPU space

When K is large, use student –teacher model

$$\mathcal{F}_t \leftarrow m\mathcal{F}_t + (1-m)\mathcal{F}_s$$

When K is small, use Temporal ensemble strategy

$$\mathbf{z}_t \leftarrow m\mathbf{z}_{t-1} + (1-m)\mathbf{z}_t$$

Algorithm 1: SimMatch (Student-Teacher)

Input: $\mathbf{x}_l$ and :$\mathbf{x}_u$ a batch of labeled and unlabeled samples. $T_w(\cdot)$ and $T_s(\cdot)$: Weak and strong augmentation function. $\mathcal{F}_t$ and $\mathcal{F}_s$: Teacher and student encoder. ϕ_t and ϕ_s: teacher and student classifier. g_t and g_s: teacher and student projection head. Q_f and Q_l: The feature and label memory buffer.

while *network not converge* **do**

  for *i=1 to step* **do** Every Batch

$\mathbf{h}^w = \mathcal{F}_t(T_w(\mathbf{x}_u))$ $\quad$ $\mathbf{h}^s = \mathcal{F}_s(T_s(\mathbf{x}_u))$

$p^w = DA(\phi_t(\mathbf{h}^w))$ $\quad$ $p^s = \phi_s(\mathbf{h}^s)$

$\mathbf{z}^w = g_t(\mathbf{h}^w)$ $\quad$ $\mathbf{z}^s = g_s(\mathbf{h}^s)$

$\mathbf{h}_t^l = \mathcal{F}_t(T_w(\mathbf{x}_l))$ $\quad$ $\mathbf{h}_s^l = \mathcal{F}_s(T_w(\mathbf{x}_l))$

$p^l = \phi_s(\mathbf{h}_s^l)$ $\quad$ $\mathbf{z}^l = g_t(\mathbf{h}_t^l)$

Compute q^w and q^s by Eq.(3) Eq.(4)

Compute p^{unfold} and q^{agg} by Eq.(7) Eq.(9)

Compute $\widehat{q}$ and $\widehat{p}$ by Eq.(8) Eq.(10)

$\mathcal{L}_s = \frac{1}{B}\sum \mathrm{H}(y, p^l)$

$\mathcal{L}_u = \frac{1}{\mu B}\sum \mathbb{1}(\max \widehat{p} > \tau)\mathrm{H}(\widehat{p}, p^s)$

$\mathcal{L}_{in} = \frac{1}{\mu B}\sum \mathrm{H}(\widehat{q}, q^s)$

$\mathcal{L}_{overall} = \mathcal{L}_s + \lambda_u\mathcal{L}_u + \lambda_{in}\mathcal{L}_{in}$

Optimize F_s, g_s and ϕ_s by $\mathcal{L}_{overall}$

Momentum Update F_t, g_t and ϕ_t

Update Q_f and Q_l with $\mathbf{z}^l$ and y

  end

end

Output: The well trained model $\mathcal{F}_s$ and g_s

Experiments

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

Performance on CIFAR-10 and CIFAR-100:

Table 1. Top-1 Accuracy comparison (mean and std over 5 runs) on CIFAR-10 and CIFAR-100 with varying labeled set sizes.

	CIFAR-10			CIFAR-100		
Method	40 labels	250 labels	4000 labels	400 labels	2500 labels	10000 labels
Π-Model [32]	-	45.74±3.97	58.99±0.38	-	42.75±0.48	62.12±0.11
Pseudo-Labeling [33]	-	50.22±0.43	83.91±0.28	-	42.62±0.46	63.79±0.19
Mean Teacher [50]	-	67.68±2.30	90.81±0.19	-	46.09±0.57	64.17±0.24
UDA [54]	70.95±5.93	91.18±1.08	95.12±0.18	40.72±0.88	66.87±0.22	75.50±0.25
MixMatch [6]	52.46±11.50	88.95±0.86	93.58±0.10	32.39±1.32	60.06±0.37	71.69±0.33
ReMixMatch [5]	80.90±9.64	94.56±0.05	95.28±0.13	55.72±2.06	72.57±0.31	76.97±0.56
FixMatch(RA) [46]	86.19±3.37	94.93±0.65	95.74±0.05	51.15±1.75	71.71±0.11	77.40±0.12
Dash [55]	86.78±3.75	**95.44**±0.13	95.92±0.06	55.24±0.96	72.82±0.21	78.03±0.14
CoMatch [34]	93.09±1.39	95.09±0.33	-	-	-	-
SimMatch(Ours)	**94.40**±1.37	95.16±0.39	**96.04**±0.01	**62.19**±2.21	**74.93**±0.32	**79.42**±0.11

Experiments

Table 2. Experimental results on ImageNet with 1% and 10% labeled examples.

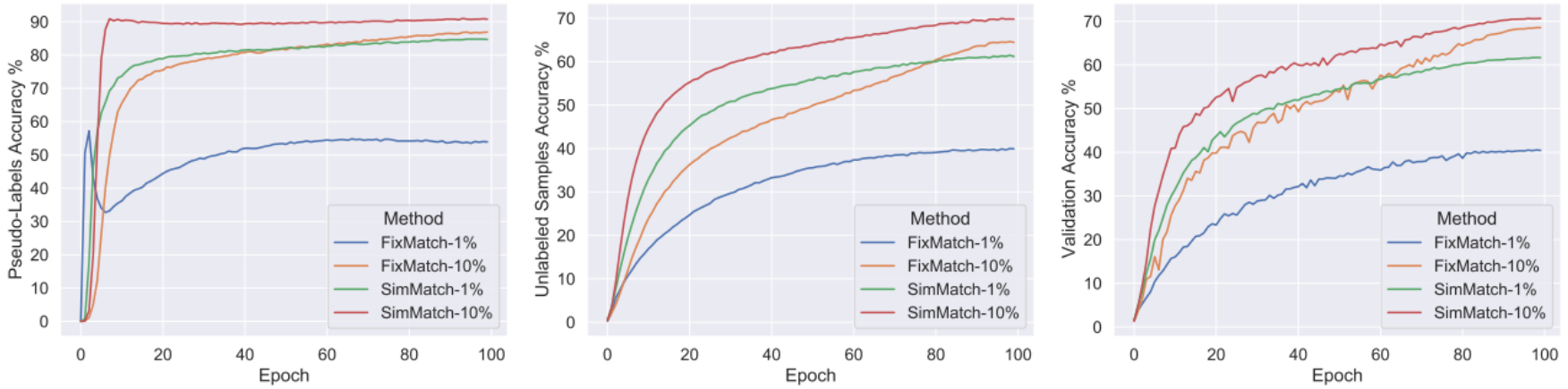
Self-supervised Pre-training	Method	Epochs	Paramters (train/test)	Top-1 Label fraction 1%	Top-1 Label fraction 10%	Top-5 Label fraction 1%	Top-5 Label fraction 10%
None	Pseudo-label [33, 58]	∼100	25.6M / 25.6M	-	-	51.6	82.4
	VAT+EntMin. [23, 39, 58]	-	25.6M / 25.6M	-	68.8	-	88.5
	S4L-Rotation [58]	∼200	25.6M / 25.6M	-	53.4	-	83.8
	UDA [54]	-	25.6M / 25.6M	-	68.8	-	88.5
	FixMatch [46]	∼300	25.6M / 25.6M	-	71.5	-	89.1
	CoMatch [34]	∼400	30.0M / 25.6M	66.0	73.6	86.4	**91.6**
PCL [35]	Fine-tune	∼200	25.8M / 25.6M	-	-	75.3	85.6
SimCLR [14]		∼1000	30.0M / 25.6M	48.3	65.6	75.5	87.8
SimCLR V2 [15]		∼800	34.2M / 25.6M	57.9	68.4	82.5	89.2
BYOL [24]		∼1000	37.1M / 25.6M	53.2	68.8	78.4	89.0
SwAV [10]		∼800	30.4M / 25.6M	53.9	70.2	78.5	89.9
WCL [61]		∼800	34.2M / 25.6M	65.0	72.0	86.3	91.2
MoCo V2 [16]	Fine-tune	∼800	30.0M / 25.6M	49.8	66.1	77.2	87.9
	CoMatch [34]	∼1200	30.0M / 25.6M	67.1	73.7	**87.1**	91.4
MoCo-EMAN [8]	FixMatch-EMAN [8]	∼1100	30.0M / 25.6M	63.0	74.0	83.4	90.9
None	**SimMatch (Ours)**	∼400	30.0M / 25.6M	**67.2**	**74.4**	**87.1**	**91.6**

Table 3. Transfer learning performance using ResNet-50 pretrained with ImageNet. Following the evaluatiion protocal from [14, 24], we report Top-1 classification accuracy except Pets and Flowers for which we report mean per-class accuracy.

Method	Epochs	CIFAR-10	CIFAR-100	Food-101	Cars	DTD	Pets	Flowers	Mean
Supervised	-	**93.6**	78.3	72.3	66.7	74.9	91.5	94.7	81.7
SimCLR [14]	1000	90.5	74.4	72.8	49.3	**75.7**	84.6	92.6	77.1
MoCo v2 [16]	800	92.2	74.6	72.5	50.5	74.4	84.6	90.5	77.0
BYOL [24]	1000	91.3	**78.4**	**75.3**	67.8	75.5	90.4	**96.1**	**82.1**
SimMatch (10%)	400	**93.6**	**78.4**	71.7	**69.7**	75.1	**92.8**	93.2	**82.1**

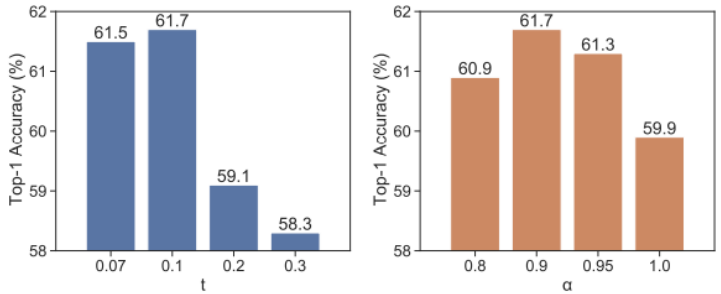
Experiments

Ablation Study:



(a) Pseudo-Labels Accuracy (b) Unlabeled Samples Accuracy (c) Validation Accuracy

Figure 4. Visualization of (a) pseudo-labels accuracy - the accuracy of $\widehat{p}$ that has higher confidence than threshold, (b) unlabeled samples accuracy - the accuracy of all $\widehat{p}$ regardless of the threshold, (c) validation accuracy for FixMatch and SimMatch on 1% and 10% setting.



(a) Temperature (b) Smooth

Figure 5. Results of varying t and α. (ImageNet-1k 1% - 100 ep)

Table 5. Results of removing scaling and smoothing strategy. (ImageNet-1k 1% - 100 ep)

Method	w/o $\widehat{p}$	w/o $\widehat{q}$	Standard
Top-1	59.9	52.3	**61.7**

Thanks