MEGVII 旷视

Boosting Semi-Supervised Learning by Exploiting All Unlabeled Data

Yuhao Chen, Xin Tan, Borui Zhao, Zhaowei Chen, Renjie Song, Jiajun Liang, Xuequan Lu

Paper tag: TUE-PM-327

Quick Preview

For FixMatch-based methods, a high confidence threshold τ plays a role of filtering noise ones, but also leading to low data utilization (blue curves)

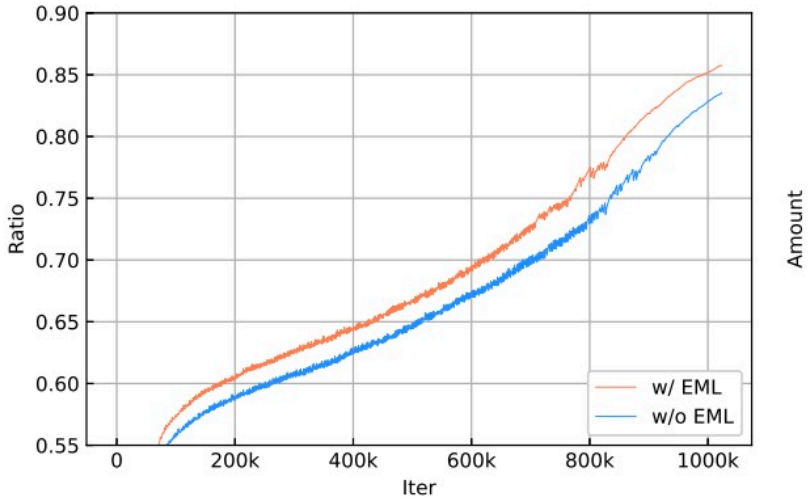
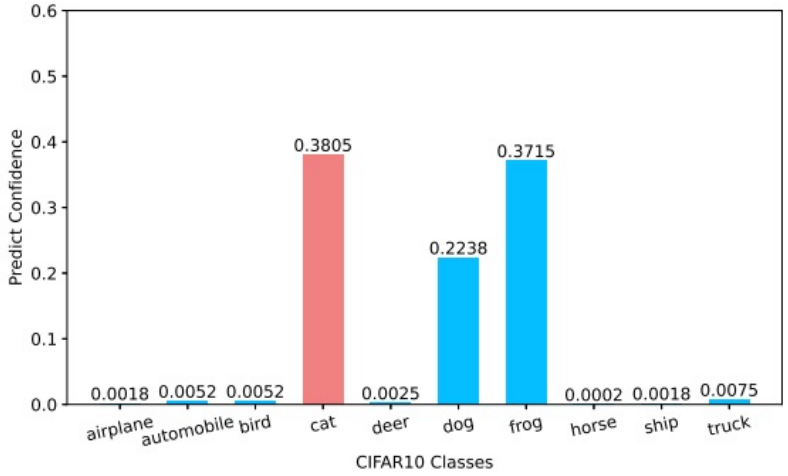


Figure 1. Visualization of the experimental results on CIFAR-100 with 10000 labeled data.

how to tackle this issue ?

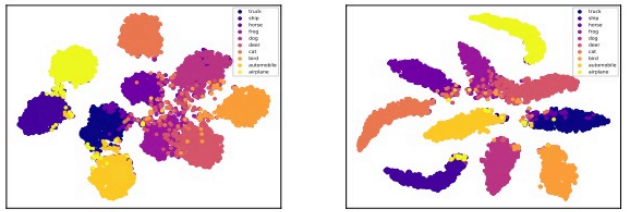
- Assign pseudo-label for potential example ?
  - make more low-entropy prediction
- how to use low-confident data (mostly wrong pseudo-label)?
  - Introduce negative pseudo-label



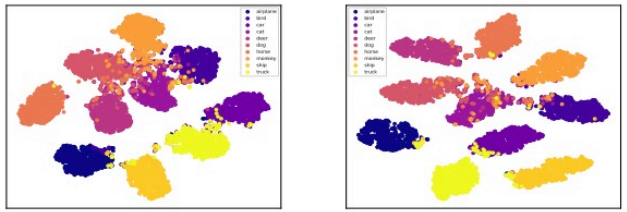
Quick Preview

Entropy Meaning Loss (EML)

EML impose an additional constraint on all non-target classes, aims to share the remaining confidence $(1 - p_{tc})$ equaly.

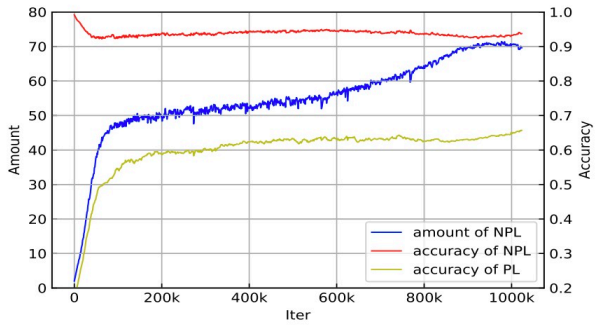


(a) CIFAR-10



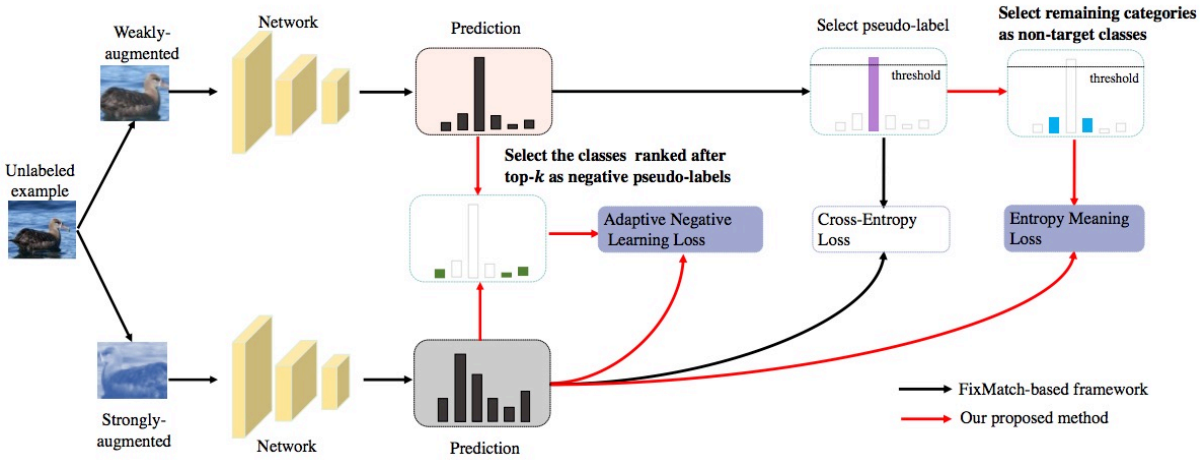
(b) STL-10

Adaptive Negative Learning (ANL)



ANL dynamic estimate a suitable k so that model's *top-k* accuracy is closed to 100%, and then categories ranked after *top-k* can be assigned negative pseudo-labels.

Our proposed FullMatch Framework



Quick Preview

- FullMatch and FullFlex are both effective, and achieving the SOTA performance in CIFAR-10/100, SVHN and STL-10.

Label Amount	CIFAR-10 40	CIFAR-10 250	CIFAR-10 4000	CIFAR-100 400	CIFAR-100 2500	CIFAR-100 10000	SVHN 40	SVHN 1000	STL-10 1000
UDA [33]	89.38±3.75	94.84±0.06	95.71±0.07	53.61±1.59	72.27±0.21	77.51±0.23	94.88±4.27	**98.11**±0.01	93.36±0.17
RemixMatch [2]	90.12±1.03	93.7±0.05	95.16±0.01	57.25±1.05	73.97±0.35	**79.98**±0.27	75.96±9.13	94.84±0.31	93.26±0.14
Semco‡ [21]	92.13±0.22	94.88±0.27	96.20±0.08	55.89±1.18	68.07±0.01	75.55±0.12	-	-	92.51±0.29
Dash [34]	86.78±3.75	95.44±0.13	95.92±0.06	55.24±0.96	72.82±0.21	78.03±0.14	96.97±1.59	97.97±0.06	92.74±0.40
UPS [24]	94.74±0.29	94.89±0.08	95.75±0.05	58.93±1.66	72.86±0.24	78.03±0.23	-	-	93.98±0.28
AlphaMatch‡ [11]	91.35±3.38	95.03±0.29	-	61.26±3.13†	**74.98**±0.27†	-	97.03±0.26	-	90.36±0.75
CoMatch [18]	93.12±0.92	95.10±0.35	95.94±0.03	59.98±1.11	72.99±0.21	78.17±0.23	-	-	91.34±0.41
SimMatch†‡ [41]	94.40±1.37	95.16±0.39	96.04±0.01	62.19±2.21	74.93±0.32	79.42±0.11	-	-	-
CR [9]	94.31±0.9	94.96±0.3	95.84±0.13	50.77±0.79	72.42±0.37	78.97±0.23	96.33±1.84	97.61±0.06	93.04±0.42
NP-Match [30]	95.09±0.04	95.04±0.06	95.89±0.02	61.08±0.99	73.97±0.26	78.78±0.13	-	-	94.41±0.24
FixMatch [26]	92.53±0.28	95.14±0.05	95.79±0.08	57.45±1.76	71.97±0.16	77.8±0.12	96.19±1.18	**98.04**±0.03	93.75±0.33
FullMatch (ours)	**94.11**±1.01	**95.36**±0.12	**96.25**±0.08	**59.42**±1.40	**73.06**±0.40	**78.56**±0.10	**97.65**±0.10	98.01±0.03	**94.26**±0.09
FlexMatch [38]	95.03±0.06	95.02±0.09	95.81±0.01	60.06±1.62	73.51±0.2	78.1±0.15	96.08±1.24	97.37±0.06	94.23±0.18
FullFlex (ours)	**95.56**±0.15	**95.61**±0.04	**96.28**±0.03	**62.60**±0.64	**74.60**±0.42	**79.26**±0.21	**97.48**±0.04	**97.58**±0.02	**94.50**±0.12

	Top-1	Top-5
UPS [24]	57.31	79.77
NP-Match [30]	58.22	80.67
FixMatch [26]	56.34	78.20
FullMatch (ours)	**57.44** (+1.1)	**79.26** (+1.06)
FlexMatch [38]	58.15	80.52
FullFlex (ours)	**59.58** (+1.43)	**81.38** (+0.86)

- Following TorchSSL, we conduct experiments on ImageNet, obviously our method are also effective.

	CE	BCE	w PL	w/o PL	Accuracy	Δ
FixMatch					57.68	-
EML	✓				58.35	+0.67
		✓			**58.47**	+0.79
ANL			✓		57.83	+0.15
				✓	58.59	+0.91
			✓	✓	**58.67**	+0.99
FullMatch		✓	✓	✓	**59.32**	+1.64

- Ablation Study on CIFAR-100. EML and ANL are all useful and can promote each other.

Motivation

FixMatch (NeurIPS 2020)

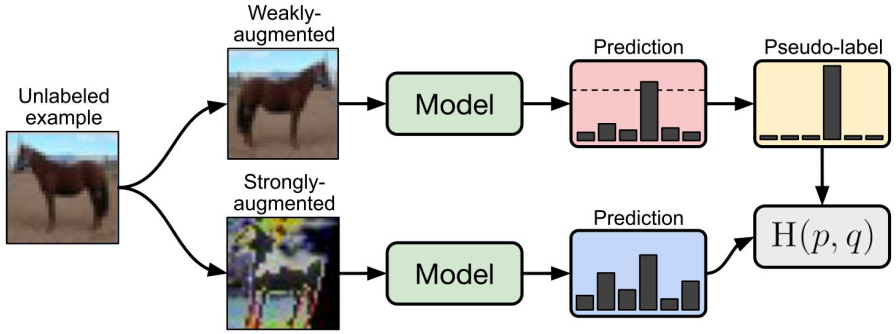


Figure 1: Diagram of FixMatch. A weakly-augmented image (top) is fed into the model to obtain predictions (red box). When the model assigns a probability to any class which is above a threshold (dotted line), the prediction is converted to a one-hot pseudo-label. Then, we compute the model's prediction 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.

For FixMatch-based methods, a high confidence threshold τ plays a role of filtering noise ones, but also leading to low data utilization (blue curves)

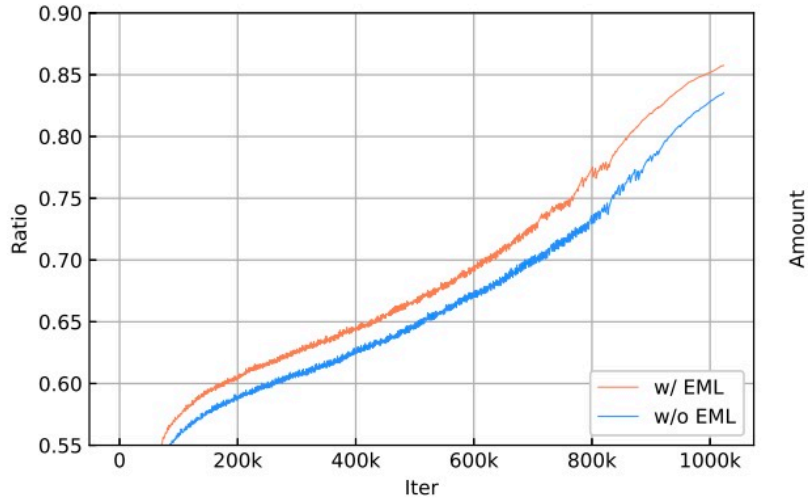


Figure 1. Visualization of the experimental results on CIFAR-100 with 10000 labeled data.

[1] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raf- fel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi- supervised learning with consistency and confidence. NeurIPS, 33, 2020

Motivation

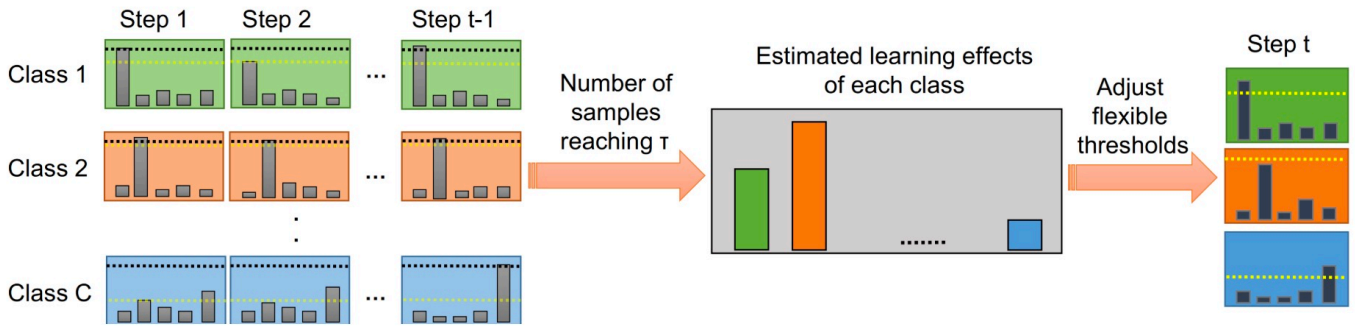
Dynamic Threshold Methods based on FixMatch

Dash (ICML 2021)

Algorithm 1 Dash: Semi-Supervised Learning with **D**yna**m**ic Thre**sh**olding

Input: learning rate η_0 and mini-batch size m_0 for stage one, learning rate η and parameter m of mini-batch size for stage two, two parameters $C > 1$ and $\gamma > 1$ for computing threshold, and violation probability δ.
// Warm-up Stage: run SGD in T_0 iterations.
Initialization: $\mathbf{u}_0 = \mathbf{w}_0$
for $t = 0, 1, \ldots, T_0 - 1$ **do**
  Sample m_0 examples $\xi_{t,i}$ $(i = 1, \ldots, m_0)$ from $\mathbf{D}_l$,
  $\mathbf{u}_{t+1} = \mathbf{u}_t - \eta_0 \tilde{\mathbf{g}}_t$ where $\tilde{\mathbf{g}}_t = \frac{1}{m_0} \sum_{i=1}^{m_0} \nabla f_s(\mathbf{u}_t; \xi_{t,i})$
end for
// Selection Stage: run SGD in T iterations.
Initialization: $\mathbf{w}_1 = \mathbf{u}_{T_0}$.
Compute the value of $\widehat{\rho}$ as in (16). // In practice, $\widehat{\rho}$ can be obtained as in (17).
for $t = 1, \ldots, T$ **do**
  1) Sample $n_t = m\gamma^{t-1}$ examples from $\mathbf{D}_u$, where the pseudo labels in $\mathbf{D}_u$ are generated by FixMatch
  2) Set the threshold $\rho_t = C\gamma^{-(t-1)}\widehat{\rho}$.
  3) Compute truncated stochastic gradient $\mathbf{g}_t$ as (18).
  4) Update solution by SGD using stochastic gradient $\mathbf{g}_t$ and learning rate η: $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{g}_t$.
end for
Output: $\mathbf{w}_{T+1}$

FlexMatch (NeurIPS 2021)



Limitations of Prior Works

- Low threshold risks introducing noise;
- Data wastes still exists, i.e., low-confident data are hard to involve to model training.

Introduction

For potential example (the maximum confidence is close to the predefined threshold), We propose Entropy Meaning Loss (EML).

- EML imposes additional supervision on all non-target to push their prediction close to a uniform distribution, thus preventing any class competition with the target class.
- Different with previous works, EML aims to make more low-entropy prediction.
- Benefit from untuning the threshold, EML can be also applied to any dynamic-threshold methods.

Introduction

For low-confident example (the maximum confidence is far from to the predefined threshold),
We propose Adaptive Negative Learning (ANL).

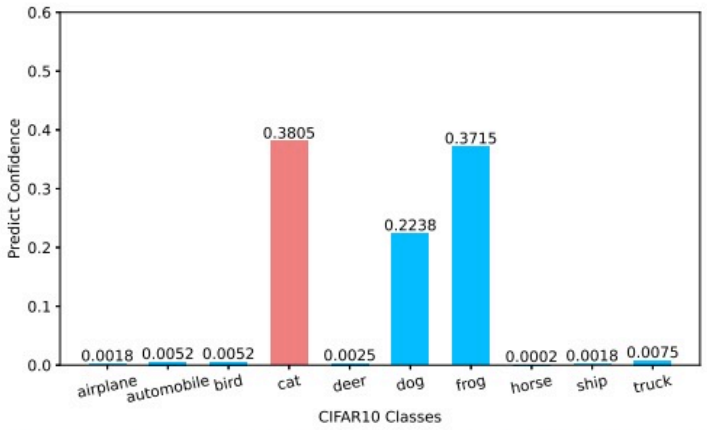


Figure 2. An example of inference result of FixMatch. It can conclude that the input does not belong to these low-rank classes, such as airplanes, horse.

ANL dynamic calculate a *k* so that the accuracy of *top-k* is close to 1, and then regard the classes ranked after *k* as negative pseudo-labels

FixMatch is confused by several top classes (e.g., “dog”, “frog”), however it shows highly confidence that some low-rank classes (e.g.,“airplane”, “horse”) are not ground truth class.

Method

Entropy Meaning Loss (EML)

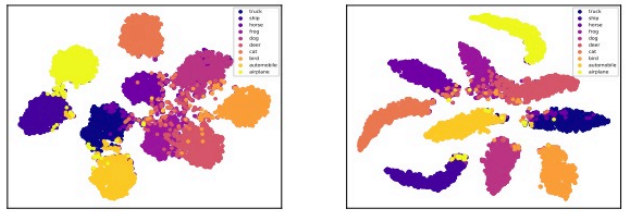
EML impose an additional constraint on all non-target classes, aims to share the remaining confidence equaly.

$$s_c^{(i)} = \mathbb{1}(q_c^{(i)} \geq \tau)$$

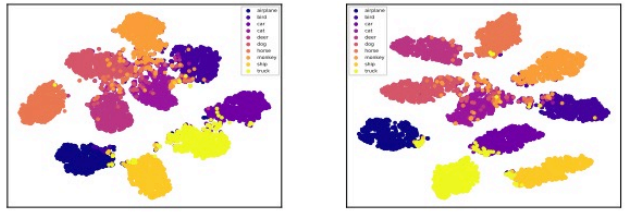
$$u_c^{(i)} = \mathbb{1}(max(Q^{(i)}) \geq \tau) \cdot \mathbb{1}(s_c^{(i)} = 0)$$

$$y_c^{(i)} = \frac{1 - \mathbb{1}(u_c^{(i)} = 0) \cdot p_c^{(i)}}{\sum_c \mathbb{1}(u_c^{(i)} = 1)}$$

$$\mathcal{L}_{eml} = -\frac{1}{BC} \sum_{i=1}^{B} \sum_{c=1}^{C} u_c^{(i)} \cdot [y_c^{(i)} log(p_c^{(i)}) + (1 - y_c^{(i)})(log(1 - p_c^{(i)}))]$$



(a) CIFAR-10



(b) STL-10

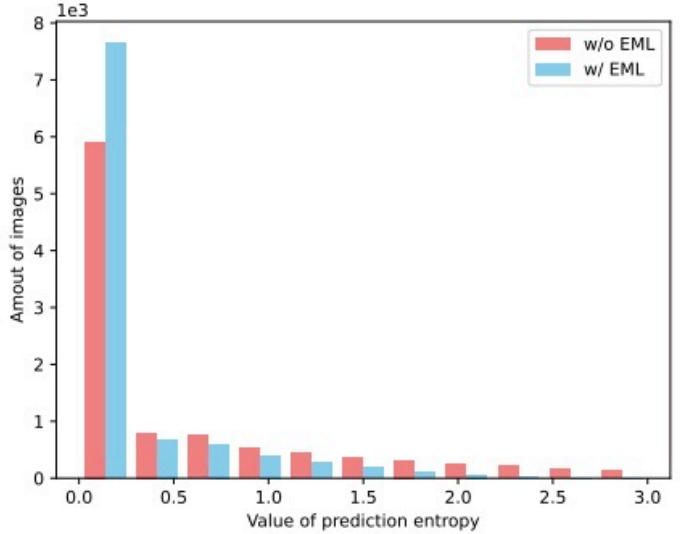


Figure 3. Visualization the distribution of prediction entropy when adopting EML to FixMatch on CIFAR-100 testset . The model supervised by EML can generate more low-entropy predictions and thus select more examples with pseudo-label.

Method

Adaptive Negative Learning (ANL)

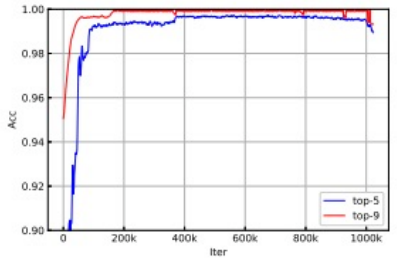


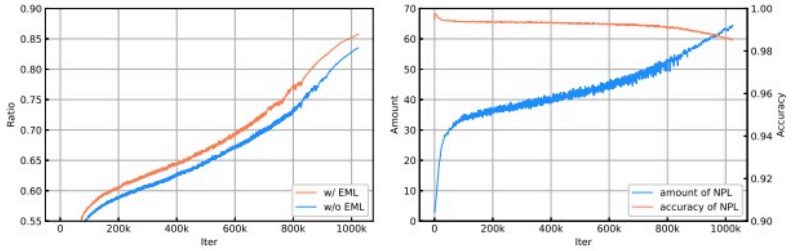
Figure 6. The top-5 and top-9 accuracy curves of FixMatch during training on CIFAR-10 with 40 label samples.

- calculate *temp* labels based weakly-aug predict;
- calculate minimum k to top-k acc is 100% based strongly-aug and *temp* labels;

$$k = \arg\min_{\theta \in [2,C]} (Acc(P_t, \hat{Q}_t, \theta) = 100\%)$$

- Ranked after k in weakly-aug prediction are assigned negative pseudo-labels.

$$\mathcal{L}_{anl} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{C} \mathbb{1}[Rank(q_c^{(i)}) > k]\, log(1 - p_c^{(i)})$$



(a) (b)

Figure 1. Visualization of the experimental results on CIFAR-100 with 10000 labeled images. Evaluations are done every 1K iterations. (a) The increasing proportion of examples with pseudo-label when applying EML to FixMatch. (b) The number of negative pseudo-labels per sample and accuracy during the whole training process. "NPL" denotes negative pseudo-labels.

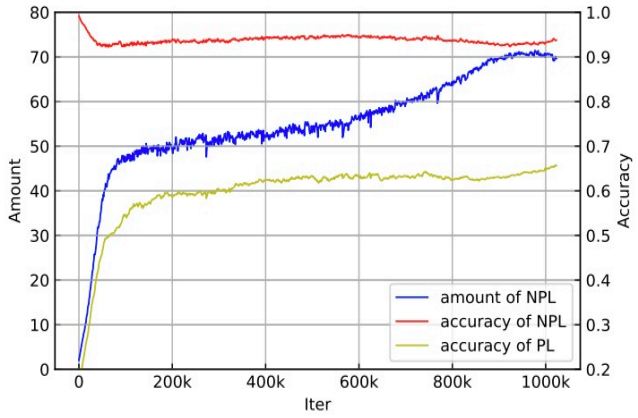


Figure 1. Visualize the experimental results on CIFAR-100 with 400 label samples.

Method

Overall Framework

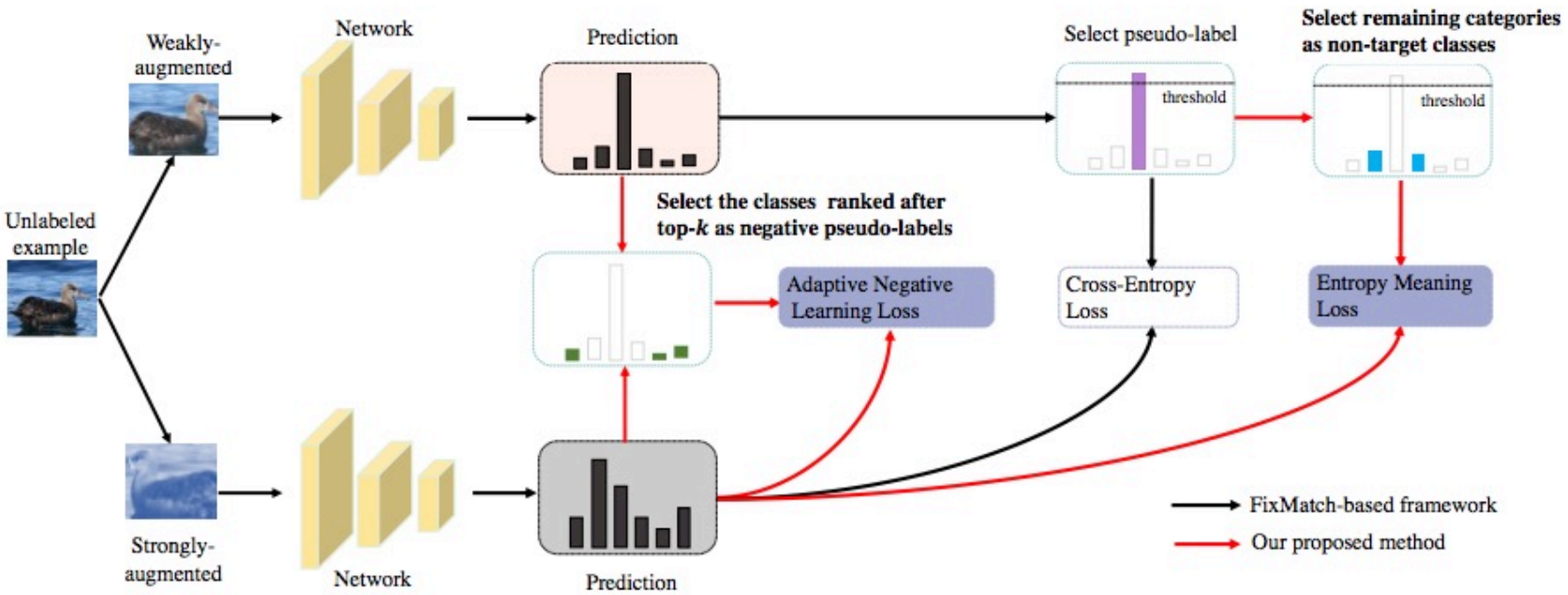


Figure 5. Overview of the proposed FullMatch. First, we allocate the negative pseudo-label (green bar) for all unlabeled data with the proposed Adaptive Negative Learning. Then, if the highest probability is above the predefined threshold (dotted line), we will assign the pseudo-label (purple bar) just like FixMatch, but we optimize further remaining non-target classes (blue bar) via the proposed Entropy Meaning Loss. The black line indicates the existing FixMatch-based methods, and the red line is our proposed method. (Best viewed in color).

Experiments

Main Results in four popular SSL benchmarks

	CIFAR-10			CIFAR-100			SVHN		STL-10
Label Amount	40	250	4000	400	2500	10000	40	1000	1000
UDA [33]	89.38±3.75	94.84±0.06	95.71±0.07	53.61±1.59	72.27±0.21	77.51±0.23	94.88±4.27	**98.11**±0.01	93.36±0.17
RemixMatch [2]	90.12±1.03	93.7±0.05	95.16±0.01	57.25±1.05	73.97±0.35	**79.98**±0.27	75.96±9.13	94.84±0.31	93.26±0.14
Semco‡ [21]	92.13±0.22	94.88±0.27	96.20±0.08	55.89±1.18	68.07±0.01	75.55±0.12	-	-	92.51±0.29
Dash [34]	86.78±3.75	95.44±0.13	95.92±0.06	55.24±0.96	72.82±0.21	78.03±0.14	96.97±1.59	97.97±0.06	92.74±0.40
UPS [24]	94.74±0.29	94.89±0.08	95.75±0.05	58.93±1.66	72.86±0.24	78.03±0.23	-	-	93.98±0.28
AlphaMatch‡ [11]	91.35±3.38	95.03±0.29	-	61.26±3.13†	**74.98**±0.27†	-	97.03±0.26	-	90.36±0.75
CoMatch [18]	93.12±0.92	95.10±0.35	95.94±0.03	59.98±1.11	72.99±0.21	78.17±0.23	-	-	91.34±0.41
SimMatch†‡ [41]	94.40±1.37	95.16±0.39	96.04±0.01	62.19±2.21	74.93±0.32	79.42±0.11	-	-	-
CR [9]	94.31±0.9	94.96±0.3	95.84±0.13	50.77±0.79	72.42±0.37	78.97±0.23	96.33±1.84	97.61±0.06	93.04±0.42
NP-Match [30]	95.09±0.04	95.04±0.06	95.89±0.02	61.08±0.99	73.97±0.26	78.78±0.13	-	-	94.41±0.24
FixMatch [26]	92.53±0.28	95.14±0.05	95.79±0.08	57.45±1.76	71.97±0.16	77.8±0.12	96.19±1.18	**98.04**±0.03	93.75±0.33
FullMatch (ours)	**94.11**±1.01	**95.36**±0.12	**96.25**±0.08	**59.42**±1.40	**73.06**±0.40	**78.56**±0.10	**97.65**±0.10	98.01±0.03	**94.26**±0.09
FlexMatch [38]	95.03±0.06	95.02±0.09	95.81±0.01	60.06±1.62	73.51±0.2	78.1±0.15	96.08±1.24	97.37±0.06	94.23±0.18
FullFlex (ours)	**95.56**±0.15	**95.61**±0.04	**96.28**±0.03	**62.60**±0.64	**74.60**±0.42	**79.26**±0.21	**97.48**±0.04	**97.58**±0.02	**94.50**±0.12

Table 1. **Top-1 accuracy (%) for CIFAR-10/100, SVHN and STL-10 datasets on 3 different folds.** *FullFlex* indicates applying our method to FlexMatch. † indicates introducing an additional technique named DA (Distribution Alignment) [2]. ‡ represents the result comes from the original paper.

	Top-1	Top-5
UPS [24]	57.31	79.77
NP-Match [30]	58.22	80.67
FixMatch [26]	56.34	78.20
FullMatch (ours)	**57.44** (+1.1)	**79.26** (+1.06)
FlexMatch [38]	58.15	80.52
FullFlex (ours)	**59.58** (+1.43)	**81.38** (+0.86)

Table 2. **Top-1 and Top-5 accuracy (%) on ImageNet.** In green are the values of performance improvement over the baselines.

Compared SOTA methods on ImageNet

Experiments

Ablation Study

	CE	BCE	w PL	w/o PL	Accuracy	Δ
FixMatch					57.68	-
EML	✓				58.35	+0.67
		✓			**58.47**	+0.79
ANL			✓		57.83	+0.15
				✓	58.59	+0.91
			✓	✓	**58.67**	+0.99
FullMatch		✓	✓	✓	**59.32**	+1.64

Table 3. **Ablation study of FullMatch on 400-label split from CIFAR-100.** CE and BCE represent the loss implementation of EML. "w PL" and "w/o PL" means applying ANL on examples with/without pseudo-label, respectively. Δ represents the performance improvement over the baseline.

Effect on different components

α	0.5			1.0			2.0		
β	0.5	1.0	2.0	0.5	1.0	2.0	0.5	1.0	2.0
acc	78.43	78.36	78.50	78.38	78.46	78.48	78.49	78.31	78.47

Table 4. **Ablation study on α and β.** All experiments are conducted on CIFAR-100 with 10000-label.

Experiment on different coefficients of EML and ANL

Conclusion

To summarize, our key contributions include:

- We propose an extra loss function, namely EML. It can help to generate more low-entropy prediction under the same threshold.
- We propose ANL, a novel negative pseudo-labels allocation scheme. It can involves all samples into model learning, including low-confidence ones.
- By integrated EML and ANL into FixMatch, we proposed FullMatch framework. It achieves remarkable gains on five benchmark.
- Our method is shown to be orthogonal to other FixMatch-based framework. Specifically, FlexMatch with our method, achieves state-of-the-art results.