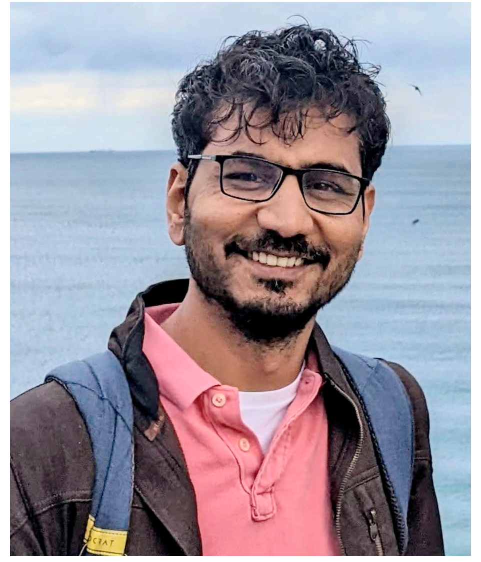
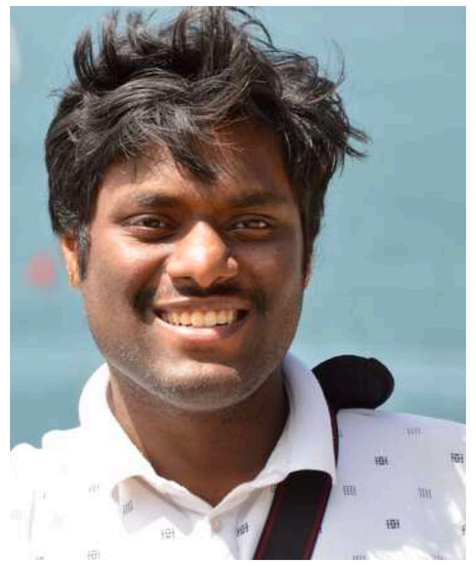
Rethinking Confidence Scores and Thresholds in Pseudolabeling-based SSL

ICML, 2025


Harit Vishwakarma[w]


Yi (Reid) Chen[w]


Srinath Namburi[g]


Sui Jiet Tay[n]


Ramya Korlakai Vinayak[w]


Frederic Sala[w]

[w] WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

[g] GE HealthCare

[n] NYU COURANT INSTITUTE OF MATHEMATICAL SCIENCES

Pseudolabeling-based Semi-supervised Learning (SSL)

Pseudolabeling-based methods (self-training) are simple, popular and actively researched.

Scudder, 1965; Blum & Mitchell, 1998; Rosenberg et al., 2005; Lee, 2013; Oymak & Gulcu, 2020; Amini et al., 2023

Pseudolabeling-based Semi-supervised Learning (SSL)

Pseudolabeling-based methods (self-training) are simple, popular and actively researched.

Scudder, 1965; Blum & Mitchell, 1998; Rosenberg et al., 2005; Lee, 2013; Oymak & Gulcu, 2020; Amini et al., 2023

Idea

Train a model $\hat{h}$ on groundtruth labeled data.

Pseudolabeling-based Semi-supervised Learning (SSL)

Pseudolabeling-based methods (self-training) are simple, popular and actively researched.

Scudder, 1965; Blum & Mitchell, 1998; Rosenberg et al., 2005; Lee, 2013; Oymak & Gulcu, 2020; Amini et al., 2023

Idea

Train a model $\hat{h}$ on groundtruth labeled data.

↓

Pseudolabel **selected** unlabeled points using $\hat{h}$.

(Selection based on **model's confidence**)

Model-predicted labels are called **pseudolabels.**

Pseudolabeling-based Semi-supervised Learning (SSL)

Pseudolabeling-based methods (self-training) are simple, popular and actively researched.

Scudder, 1965; Blum & Mitchell, 1998; Rosenberg et al., 2005; Lee, 2013; Oymak & Gulcu, 2020; Amini et al., 2023

Idea

Train a model $\hat{h}$ on groundtruth labeled data.

$\downarrow$

Pseudolabel **selected** unlabeled points using $\hat{h}$.

(Selection based on **model's confidence**)

$\downarrow$

Update $\hat{h}$ by training on the groundtruth labeled and pseudolabeled data.

Model-predicted labels are called **pseudolabels.**

Pseudolabeling-based Semi-supervised Learning (SSL)

Pseudolabeling-based methods (self-training) are simple, popular and actively researched.

Scudder, 1965; Blum & Mitchell, 1998; Rosenberg et al., 2005; Lee, 2013; Oymak & Gulcu, 2020; Amini et al., 2023

Idea

Train a model $\hat{h}$ on groundtruth labeled data.

Pseudolabel **selected** unlabeled points using $\hat{h}$.

(Selection based on **model's confidence**)

Update $\hat{h}$ by training on the groundtruth labeled and pseudolabeled data.

Model-predicted labels are called **pseudolabels.**

Pseudolabeling-based Semi-supervised Learning (SSL)

Pseudolabeling-based methods (self-training) are simple, popular and actively researched.

Scudder, 1965; Blum & Mitchell, 1998; Rosenberg et al., 2005; Lee, 2013; Oymak & Gulcu, 2020; Amini et al., 2023

Idea

Train a model $\hat{h}$ on groundtruth labeled data.

↓

Pseudolabel **selected** unlabeled points using $\hat{h}$.

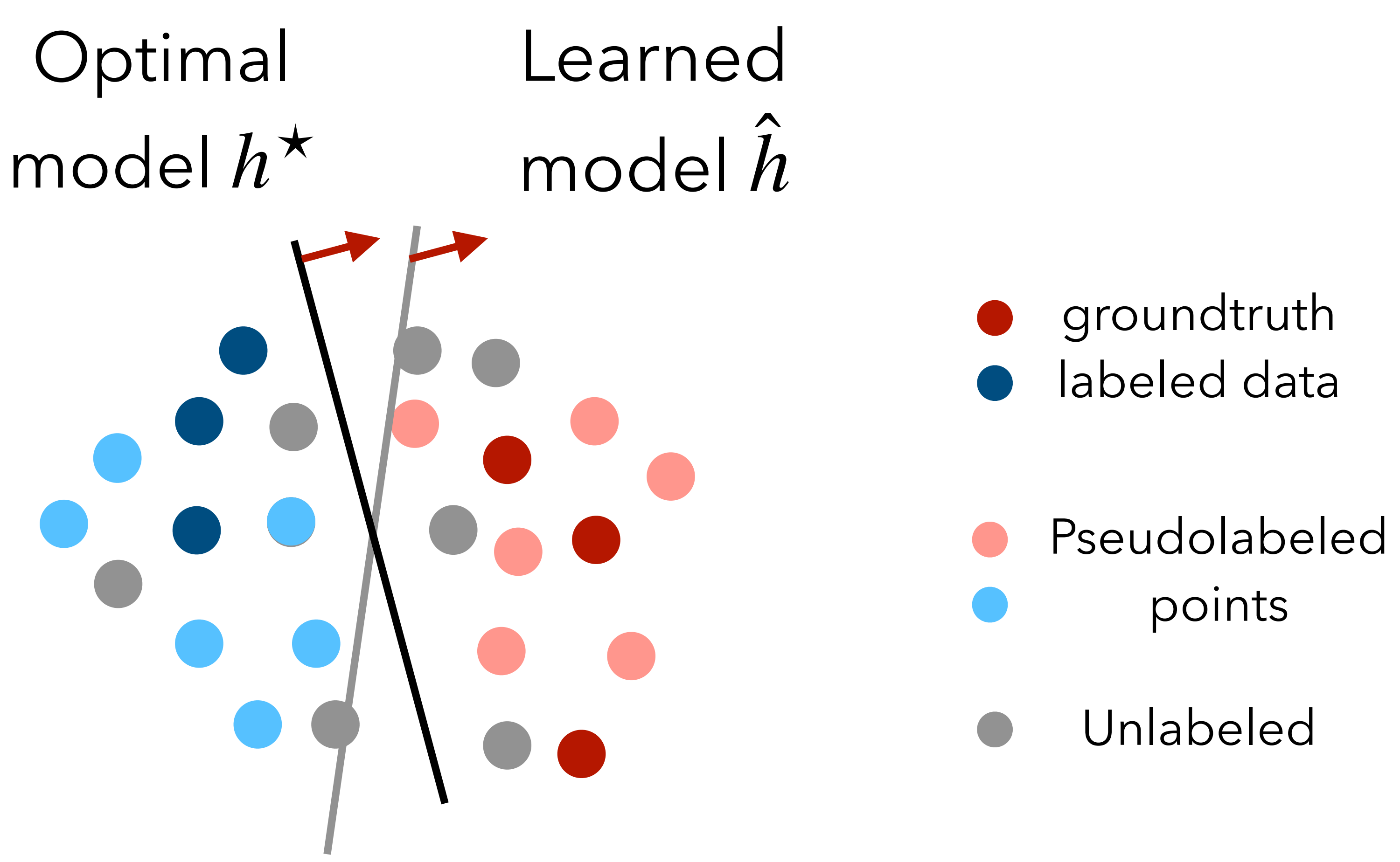
(Selection based on **model's confidence**)

↓

Update $\hat{h}$ by training on the groundtruth labeled and pseudolabeled data.

Model-predicted labels are called **pseudolabels.**

Expectation



Using groundtruth and pseudolabeled data better model can be learned.

Pseudolabeling-based Semi-supervised Learning (SSL)

Pseudolabeling-based methods (self-training) are simple, popular and actively researched.

Scudder, 1965; Blum & Mitchell, 1998; Rosenberg et al., 2005; Lee, 2013; Oymak & Gulcu, 2020; Amini et al., 2023

Idea

Train a model $\hat{h}$ on groundtruth labeled data.

Pseudolabel **selected** unlabeled points using $\hat{h}$.

(Selection based on **model's confidence**)

Update $\hat{h}$ by training on the groundtruth labeled and pseudolabeled data.

Quality and quantity of pseudolabeled data

Learning curve (speed)
Test accuracy of the end model.

Model-predicted labels are called **pseudolabels.**

Common Selection (Pseudolabeling) Strategy

Pseudolabel points having **confidence score** above a certain **threshold**.

Common Selection (Pseudolabeling) Strategy

Pseudolabel points having **confidence score** above a certain **threshold**.

Confidence Function $g : \mathcal{X} \to \Delta^k$

k: classes.

Confidence in predictions of the classifier

Depends on h but drop it for convenience

Predicted label/class

$$\hat{y} = \hat{h}(\mathbf{x})$$

Confidence Score

$$g(\mathbf{x})[\hat{y}]$$

Softmax Score

Multi-class setting

0	1	2	3
0.02	0.06	0.02	**0.9**

$\hat{y} = 3$ $\quad g(\mathbf{x})[\hat{y}] = 0.9$

Common Selection (Pseudolabeling) Strategy

Pseudolabel points having **confidence score** above a certain **threshold**.

Confidence Function $g : \mathcal{X} \to \Delta^k$

k: classes.

Confidence in predictions of the classifier

Depends on h but drop it for convenience

Predicted label/class

$$\hat{y} = \hat{h}(\mathbf{x})$$

Confidence Score

$$g(\mathbf{x})[\hat{y}]$$

Softmax Score

Multi-class setting

0.02	0.06	0.02	**0.9**
0	1	2	3

$\hat{y} = 3$ $\quad g(\mathbf{x})[\hat{y}] = 0.9$

$\mathbf{t} \in [0,1]^k$ Thresholds for each of the k-classes.

$\mathbf{t}[\hat{y}]$ Threshold for class $\hat{y}$

Common Selection (Pseudolabeling) Strategy

Pseudolabel points having **confidence score** above a certain **threshold**.

Confidence Function $g : \mathcal{X} \to \Delta^k$

k: classes.

Confidence in predictions of the classifier

Depends on h but drop it for convenience

Predicted label/class

$$\hat{y} = \hat{h}(\mathbf{x})$$

Confidence Score

$$g(\mathbf{x})[\hat{y}]$$

Softmax Score

Multi-class setting

0	1	2	3
0.02	0.06	0.02	**0.9**

$\hat{y} = 3$ $\quad g(\mathbf{x})[\hat{y}] = 0.9$

$\mathbf{t} \in [0,1]^k$ Thresholds for each of the k-classes.

$\mathbf{t}[\hat{y}]$ Threshold for class $\hat{y}$

Single global threshold t

$$\mathbf{t}[\hat{y}] = t \quad \forall \hat{y} \in \mathcal{Y}$$

Class-wise thresholds

$$\mathbf{t} \in \Delta^k$$

Pseudolabeling Coverage and Error

Selection Function

$$S(\mathbf{x}, g, \mathbf{t} \mid \hat{h}) = \mathbb{1}(g(\mathbf{x})[\hat{y}] \geq \mathbf{t}[\hat{y}])$$

Pseudolabeling Coverage and Error

Selection Function

$$S(\mathbf{x}, g, \mathbf{t} \mid \hat{h}) = \mathbb{1}(g(\mathbf{x})[\hat{y}] \geq \mathbf{t}[\hat{y}])$$

Pseudolabeling Coverage

Fraction of selected (psuedolabeled) points

$$\widehat{\mathcal{P}}(g, \mathbf{t} \mid \hat{h}, D) = \frac{1}{|D|} \sum_{(\mathbf{x}_i, y_i) \in D} S(\mathbf{x}, g, \mathbf{t} \mid \hat{h})$$

Pseudolabeling Coverage and Error

Selection Function

$$S(\mathbf{x}, g, \mathbf{t} \mid \hat{h}) = \mathbb{1}(g(\mathbf{x})[\hat{y}] \geq \mathbf{t}[\hat{y}])$$

Pseudolabeling Coverage
Fraction of selected (psuedolabeled) points

$$\widehat{\mathcal{P}}(g, \mathbf{t} \mid \hat{h}, D) = \frac{1}{|D|} \sum_{(\mathbf{x}_i, y_i) \in D} S(\mathbf{x}, g, \mathbf{t} \mid \hat{h})$$

Pseudolabeling Error
Fraction of psuedolabeled points with incorrect label

$$\widehat{\mathcal{E}}(g, \mathbf{t} \mid \hat{h}, D) = \frac{\sum_{(\mathbf{x}_i, y_i) \in D} S(\mathbf{x}, g, \mathbf{t} \mid \hat{h}) \cdot \mathbb{1}(\hat{y}_i \neq y_i)}{\sum_{(\mathbf{x}_i, y_i) \in D} S(\mathbf{x}, g, \mathbf{t} \mid \hat{h})}$$

Prior Work and Motivation

Pseudolabeling-based methods are popular choice and actively researched.

Scudder, 1965; Blum & Mitchell, 1998; Rosenberg et al., 2005; Lee, 2013; Oymak & Gulcu, 2020; Amini et al., 2023

Train a model $\hat{h}$ on groundtruth labeled data.

↓

Pseudolabel **selected** unlabeled points using $\hat{h}$.

(Selection based on **model's confidence**)

↓

Update $\hat{h}$ by training on the groundtruth labeled and pseudolabeled data.

Prior work

Model's softmax scores

Fixed or heuristic based thresholds

e.g. Fixmatch (Sohn et al, 2020) , Flexmatch (Zhang et al. 2022), etc.

Prior Work and Motivation

Pseudolabeling-based methods are popular choice and actively researched.

Scudder, 1965; Blum & Mitchell, 1998; Rosenberg et al., 2005; Lee, 2013; Oymak & Gulcu, 2020; Amini et al., 2023

Train a model $\hat{h}$ on groundtruth labeled data.

Pseudolabel **selected** unlabeled points using $\hat{h}$.

(Selection based on **model's confidence**)

Update $\hat{h}$ by training on the groundtruth labeled and pseudolabeled data.

Prior work

Model's softmax scores

Fixed or heuristic based thresholds

e.g. Fixmatch (Sohn et al, 2020) , Flexmatch (Zhang et al. 2022), etc.

Calibrated Confidence scores

(Loh et al. 2024, Mishra et al. 2024)

Prior Work and Motivation

Pseudolabeling-based methods are popular choice and actively researched.

Scudder, 1965; Blum & Mitchell, 1998; Rosenberg et al., 2005; Lee, 2013; Oymak & Gulcu, 2020; Amini et al., 2023

Train a model $\hat{h}$ on groundtruth labeled data.

Pseudolabel **selected** unlabeled points using $\hat{h}$.

(Selection based on **model's confidence**)

Update $\hat{h}$ by training on the groundtruth labeled and pseudolabeled data.

Prior work

Model's softmax scores

Fixed or heuristic based thresholds

e.g. Fixmatch (Sohn et al, 2020) , Flexmatch (Zhang et al. 2022), etc.

Calibrated Confidence scores

(Loh et al. 2024, Mishra et al. 2024)

Low coverage or high pseudolabeling errors

Bad models or very slow learning

~Million training iterations.

Our Work

Introduce **principled** choices for **confidence scores** and **thresholding.**

Train a model $\hat{h}$ on groundtruth labeled data.

Pseudolabel **selected** unlabeled points using $\hat{h}$.

(Selection based on **model's confidence**)

Update $\hat{h}$ by training on the groundtruth labeled and pseudolabeled data.

Our work

Learnable confidence scores and thresholds maximizing coverage at target pseudolabeling error

Our Work

Introduce **principled** choices for **confidence scores** and **thresholding.**

Train a model $\hat{h}$ on groundtruth labeled data.

Pseudolabel **selected** unlabeled points using $\hat{h}$.

(Selection based on **model's confidence**)

Update $\hat{h}$ by training on the groundtruth labeled and pseudolabeled data.

Our work

Learnable confidence scores and thresholds maximizing coverage at target pseudolabeling error

Better coverage and error trade-off

Better model and lesser training iterations

Learning Confidence Scores and Thresholds

In any round, given the classifier $\widehat{h}_i$

We want to find function $\widehat{g}_i$ and thresholds $\hat{\mathbf{t}}_i$ that can,

a) Give maximum coverage
b) Ensure pseudolabeling error $\leq \epsilon$

Learning Confidence Scores and Thresholds

In any round, given the classifier $\widehat{h}_i$

We want to find function $\widehat{g}_i$ and thresholds $\hat{\mathbf{t}}_i$ that can,

a) Give maximum coverage
b) Ensure pseudolabeling error $\leq \epsilon$

We adapt the approach to learn scores and thresholds from a prior work on automated data labeling [1].

[1] Pearls from Pebbles: ImPearls from Pebbles: Improved Confidence Functions for Auto-labeling, Vishwakarma et al., NeurIPS, 2024

Learning Confidence Scores and Thresholds

In any round, given the classifier $\widehat{h}_i$

We want to find function $\widehat{g}_i$ and thresholds $\hat{\mathbf{t}}_i$ that can,

a) Give maximum coverage
b) Ensure pseudolabeling error $\leq \epsilon$

We adapt the approach to learn scores and thresholds from a prior work on automated data labeling [1].

$$\widehat{g}_i, \widehat{\mathbf{t}}_i' \in \underset{g \in \mathcal{G}, \mathbf{t} \in T^k}{\arg\min} \ -\widetilde{\mathcal{P}}(g, \mathbf{t} \mid \widehat{h}_i, D_{\text{cal}}) + \lambda \, \widetilde{\mathcal{E}}(g, \mathbf{t} \mid \widehat{h}_i, D_{\text{cal}})$$

Smooth surrogates for coverage and error.

Solve it using gradient-based methods SGD, Adam etc.

[1] Pearls from Pebbles: ImPearls from Pebbles: Improved Confidence Functions for Auto-labeling, Vishwakarma et al., NeurIPS, 2024

Learning Confidence Scores and Thresholds

In any round, given the classifier $\widehat{h}_i$

We want to find function $\widehat{g}_i$ and thresholds $\hat{\mathbf{t}}_i$ that can,

a) Give maximum coverage
b) Ensure pseudolabeling error $\leq \epsilon$

We adapt the approach to learn scores and thresholds from a prior work on automated data labeling [1].

$$\widehat{g}_i, \widehat{\mathbf{t}}'_i \in \underset{g\in\mathcal{G},\mathbf{t}\in T^k}{\arg\min} \; -\widetilde{\mathcal{P}}(g,\mathbf{t} \mid \widehat{h}_i, D_{\text{cal}}) + \lambda\, \widetilde{\mathcal{E}}(g,\mathbf{t} \mid \widehat{h}_i, D_{\text{cal}})$$

Re-estimate the thresholds $\hat{\mathbf{t}}_i$ using another part of validation data

Smooth surrogates for coverage and error.

Solve it using gradient-based methods SGD, Adam etc.

Ensures the error constraint is strictly maintained

[1] Pearls from Pebbles: ImPearls from Pebbles: Improved Confidence Functions for Auto-labeling, Vishwakarma et al., NeurIPS, 2024

Learning Confidence Scores and Thresholds

In any round, given the classifier $\widehat{h}_i$

We want to find function $\widehat{g}_i$ and thresholds $\hat{\mathbf{t}}_i$ that can,

a) Give maximum coverage
b) Ensure pseudolabeling error $\leq \epsilon$

We adapt the approach to learn scores and thresholds from a prior work on automated data labeling [1].

$$\widehat{g}_i, \widehat{\mathbf{t}}'_i \in \underset{g \in \mathcal{G}, \mathbf{t} \in T^k}{\arg\min} \; -\widetilde{\mathcal{P}}(g, \mathbf{t} \mid \widehat{h}_i, D_{\text{cal}}) + \lambda \, \widetilde{\mathcal{E}}(g, \mathbf{t} \mid \widehat{h}_i, D_{\text{cal}})$$

Re-estimate the thresholds $\hat{\mathbf{t}}_i$ using another part of validation data

Smooth surrogates for coverage and error.

Solve it using gradient-based methods SGD, Adam etc.

Ensures the error constraint is strictly maintained

Adapt existing pseudolabeling methods to use $\widehat{g}_i$ and $\hat{\mathbf{t}}_i$ in the selection function.

[1] Pearls from Pebbles: ImPearls from Pebbles: Improved Confidence Functions for Auto-labeling, Vishwakarma et al., NeurIPS, 2024

Experiments and Results

Base methods

Fixmatch (Sohn et al., 2024) and Freematch (Wang et al., 2023)

Adapt with our scores and thresholds

Experiments and Results

Base methods

Fixmatch (Sohn et al., 2024) and Freematch (Wang et al., 2023)

Adapt with our scores and thresholds

Adaptations with MR (Mishra et al., 2024) and BaM (Loh et al., 2023) designed to promote calibrated scores

Experiments and Results

Base methods

Fixmatch (Sohn et al., 2024) and Freematch (Wang et al., 2023)

Adapt with our scores and thresholds

Adaptations with MR (Mishra et al., 2024) and BaM (Loh et al., 2023) designed to promote calibrated scores

Equalize the number of training iterations to maintain parity in overall training time

Experiments and Results

Base methods

Fixmatch (Sohn et al., 2024) and Freematch (Wang et al., 2023)

Adapt with our scores and thresholds

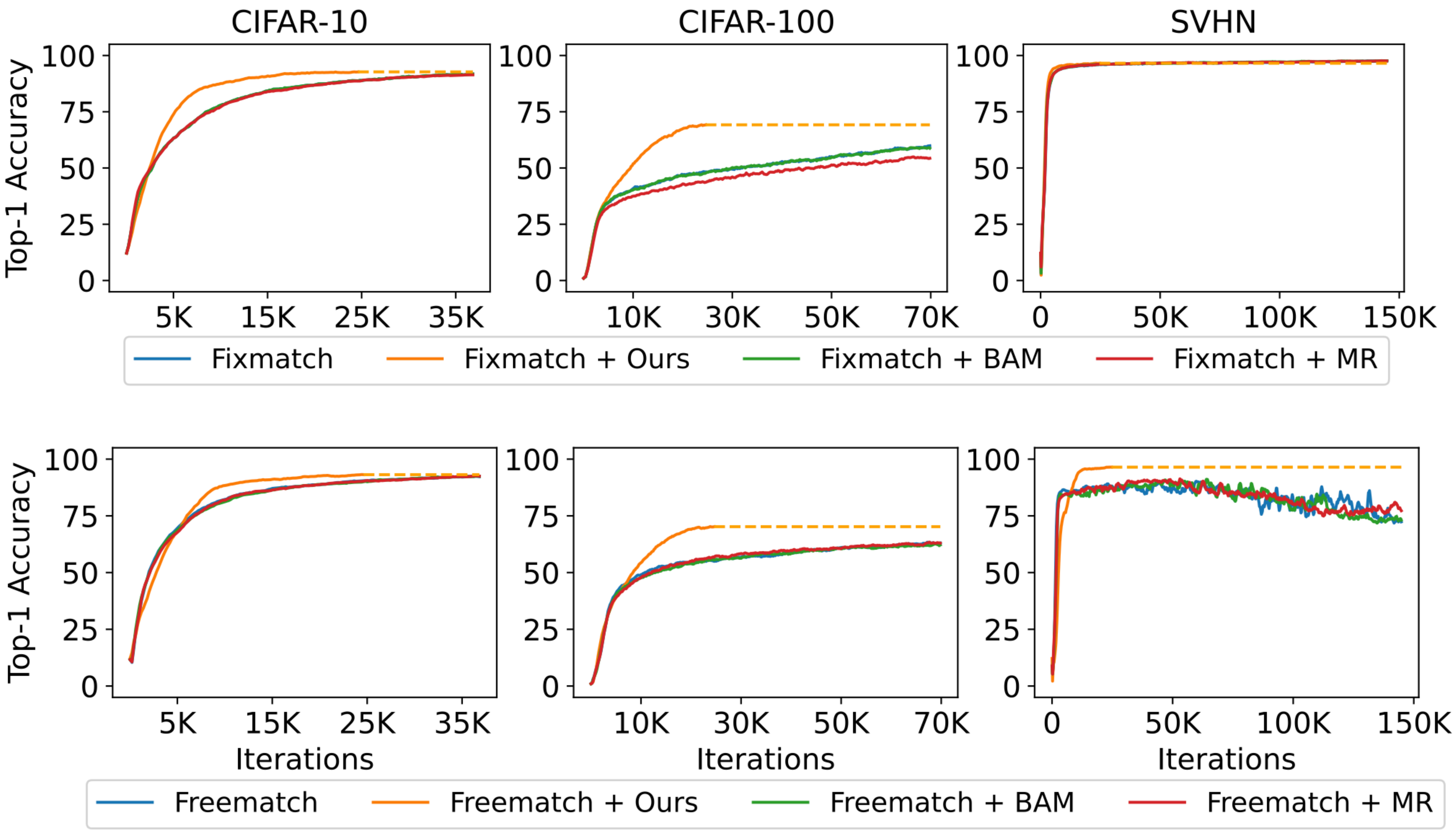
Adaptations with MR (Mishra et al., 2024) and BaM (Loh et al., 2023) designed to promote calibrated scores

Equalize the number of training iterations to maintain parity in overall training time

Dataset	CIFAR-10	CIFAR-100	SVHN
# Labels	250	2500	250
Fixmatch	90.8 ± 0.78	59.09 ± 1.10	$\mathbf{97.57} \pm \mathbf{0.08}$
Fixmatch + MR	90.41 ± 0.83	54.16 ± 0.18	97.55 ± 0.08
Fixmatch + BaM	90.67 ± 0.90	56.60 ± 2.45	97.51 ± 0.13
Fixmatch + Ours	$\mathbf{92.69} \pm \mathbf{0.74}$	$\mathbf{69.10} \pm \mathbf{0.45}$	96.54 ± 0.13
Freematch	92.26 ± 0.18	63.13 ± 0.46	92.90 ± 2.76
Freematch + MR	92.17 ± 0.36	62.03 ± 0.82	93.26 ± 2.36
Freematch + BaM	92.32 ± 0.25	62.13 ± 2.93	91.08 ± 3.72
Freematch + Ours	$\mathbf{93.10} \pm \mathbf{0.28}$	$\mathbf{68.76} \pm \mathbf{1.38}$	$\mathbf{96.65} \pm \mathbf{0.26}$

With our scores and thresholds, the base methods achieve higher accuracy at the same training time cost.

Adaptations with learned scores and thresholds help in attaining higher test accuracy earlier



Limitations and Future Work

- Our work introduced principled choices for confidence scores and thresholds in pseudolabeling-based SSL.

Limitations and Future Work

- Our work introduced principled choices for confidence scores and thresholds in pseudolabeling-based SSL.

- It relies on validation data to learn the scores and thresholds, which can be a bottleneck in practical deployments.

Limitations and Future Work

- Our work introduced principled choices for confidence scores and thresholds in pseudolabeling-based SSL.

- It relies on validation data to learn the scores and thresholds, which can be a bottleneck in practical deployments.

- Reduce the requirements of validation data

Limitations and Future Work

- Our work introduced principled choices for confidence scores and thresholds in pseudolabeling-based SSL.

- It relies on validation data to learn the scores and thresholds, which can be a bottleneck in practical deployments.

- Reduce the requirements of validation data
  - Data augmentation or generative AI to get more validation samples from small initial pool.

Limitations and Future Work

- Our work introduced principled choices for confidence scores and thresholds in pseudolabeling-based SSL.

- It relies on validation data to learn the scores and thresholds, which can be a bottleneck in practical deployments.

- Reduce the requirements of validation data
  - Data augmentation or generative AI to get more validation samples from small initial pool.
  - Carefully use the noisy pseudolabeled data to learn the scores and thresholds.

Thank You!

Contact

hvishwakarma@cs.wisc.edu

yi.chen@wisc.edu