Debiased Self-Training for Semi-Supervised Learning

Baixu Chen*, Junguang Jiang*, Ximei Wang, Pengfei Wan, Jianmin Wang, Mingsheng Long

NEURAL INFORMATION PROCESSING SYSTEMS

Overview

Research Topic

- We focus on the underexplored **bias issues in self-training**, which give rise to *training instability* and *imbalanced performance*.

Contributions

- Systematically identify the problem and **analyze the causes of self-training bias**.
- **A novel method**, Debiased-Self-Training (DST), that (1) boosts the accuracy, stability, and performance balance, and (2) can serve as a universal add-on.

Effectiveness

- DST achieves **an average boost of 6.3%** against state-of-the-art methods on standard datasets and **18.9%** against FixMatch on **13** diverse tasks.

Analysis of Bias in Self-Training

Definition

- The bias in our study refers to *deviation between the learned decision hyperplanes and the true decision hyperplanes*, measured by **the fraction of incorrectly pseudo-labeled samples in any classes**.

Causes of Self-Training Bias

- The sampling of labeled data.
- The pre-trained representations.
- The aggressive self-training strategy (e.g. FixMatch) with pseudo labels.

Figure: Effect of *labeled data sampling*.

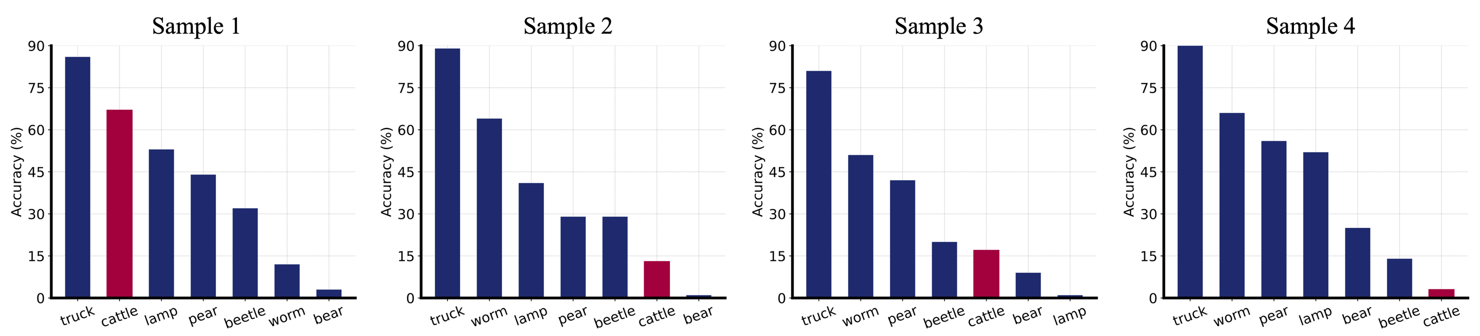
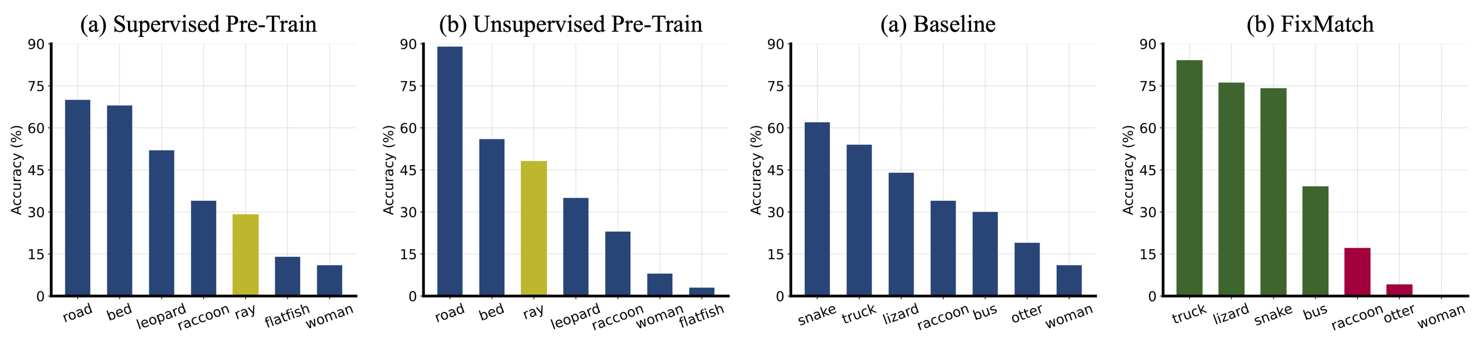
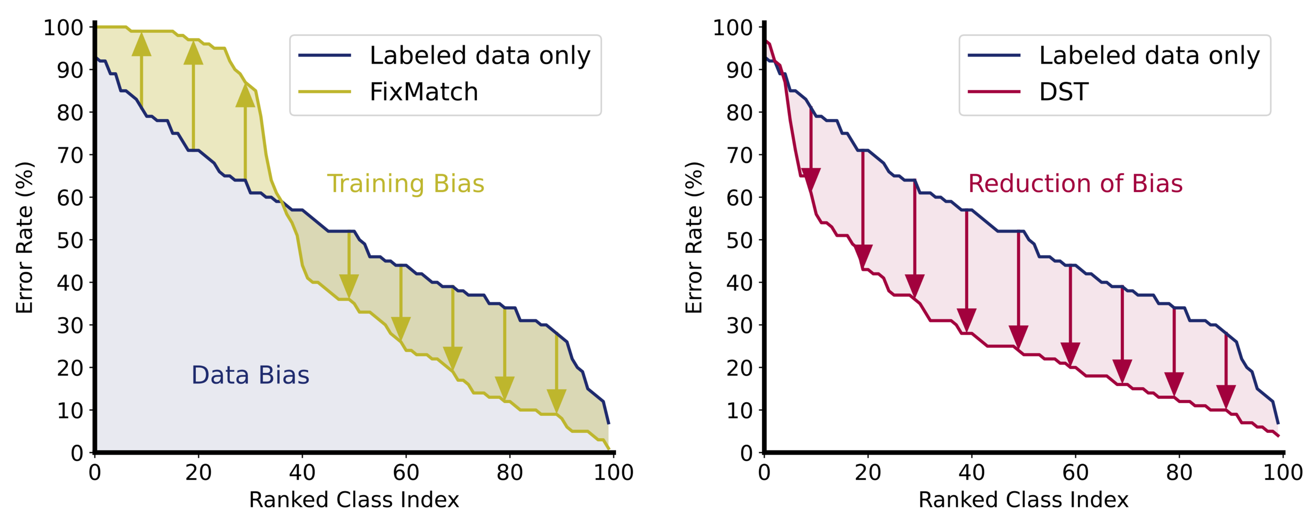


Figure: Effect of *pre-trained representations* (left) and *self-training strategy* (right).



Decomposition of Bias

- Data bias: the bias **inherent** in semi-supervised learning tasks (**blue** area), such as the bias of sampling and pre-trained representations on unlabeled data.
- Training bias: the bias **increment** brought by self-training strategies (**yellow** area).



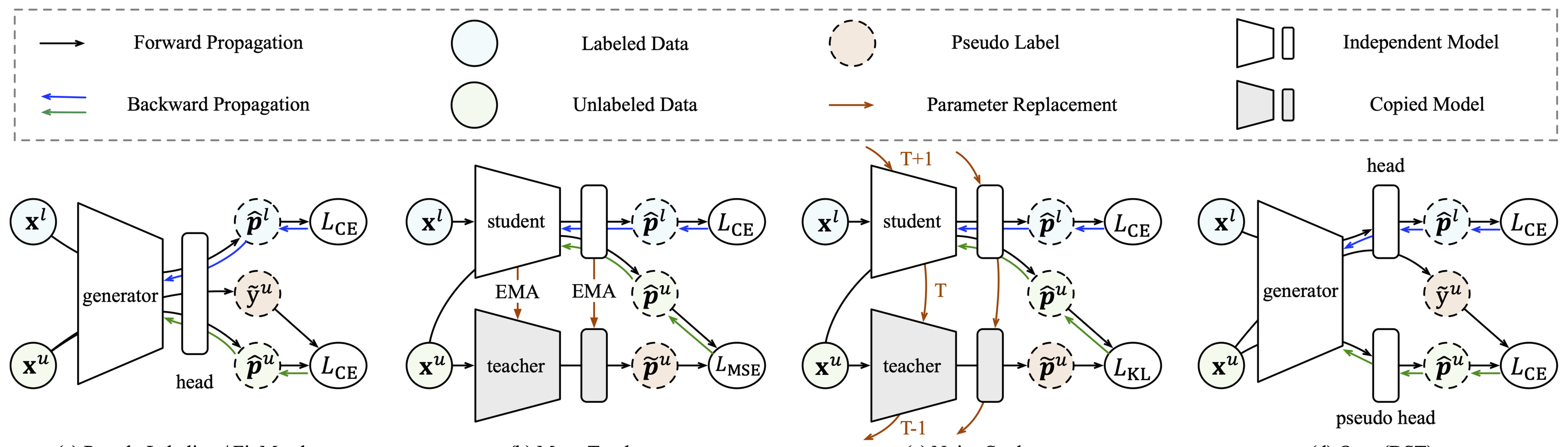
Discussion

- Data bias exists in supervised learning as well. Yes in SSL with **extreme few labeled samples**, it might cause the accuracy of the same category to vary dramatically.
- Training bias is unique in SSL and can be mitigated by better strategy.

Decrease Training Bias: Decoupled Pseudo Labeling

Insights

- Generating and utilizing pseudo labels with **the same model** amplifies bias.
- The feature generator ψ has **better tolerance** for noisy pseudo labels than the head h.



(a) Pseudo Labeling / FixMatch (b) Mean Teacher (c) Noisy Student (d) Ours (DST)

Method

- Optimize the head h only with the clean labels on labeled dataset $\mathcal{L}$ and **without** any unreliable pseudo labels from unlabeled dataset $\mathcal{U}$.
- Introduce a completely **parameter independent** pseudo head h_{pseudo}, which takes the duty of training with pseudo labels for learning a better representation.
- **The decoupled pseudo labels are generated by** h **while utilized by** h_{pseudo}.

Decrease Data Bias: Worst Case Estimation

Insights

- Training bias can be considered as the **accumulation of data bias**.
- **The worst training bias** is a good measure of data bias.

Method

- Introduce a worst possible head h', such that h' predicts perfectly on $\mathcal{L}$ while making as many mistakes as possible on $\mathcal{U}$.

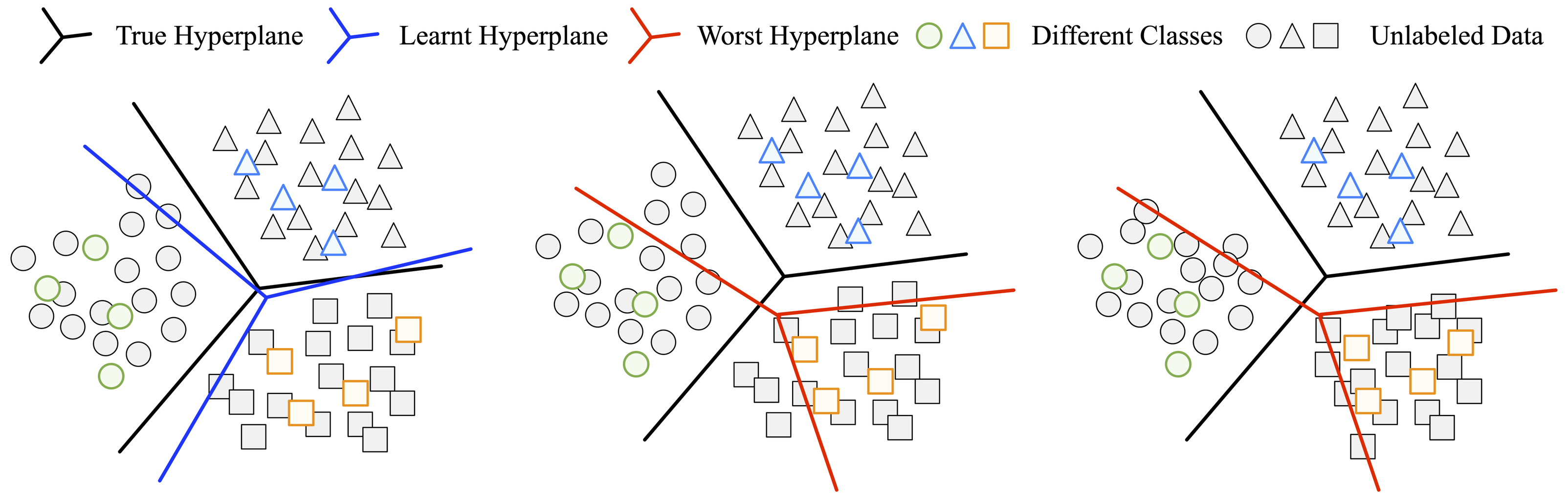
$$h_{\text{worst}}(\psi) = \arg\max_{h'} L_{\mathcal{U}}(\psi, h', \widehat{f}_{\psi,h}) - L_{\mathcal{L}}(\psi, h'). \quad (1)$$

- **Adversarially** optimize feature generator ψ to indirectly decrease the data bias.

$$\min_{\psi} L_{\mathcal{U}}(\psi, h_{\text{worst}}(\psi), \widehat{f}_{\psi,h}) - L_{\mathcal{L}}(\psi, h_{\text{worst}}(\psi)). \quad (2)$$

- Optimize ψ and h' alternatively during training, similar to GAN.

Illustration



(a) (b) (c)

- **Explanation**: **(a)** Shift between the hyperplanes learned and the true hyperplanes. **(b)** The worst hyperplanes achieved by h'. **(c)** Optimized feature representations of ψ.

Overall Objective

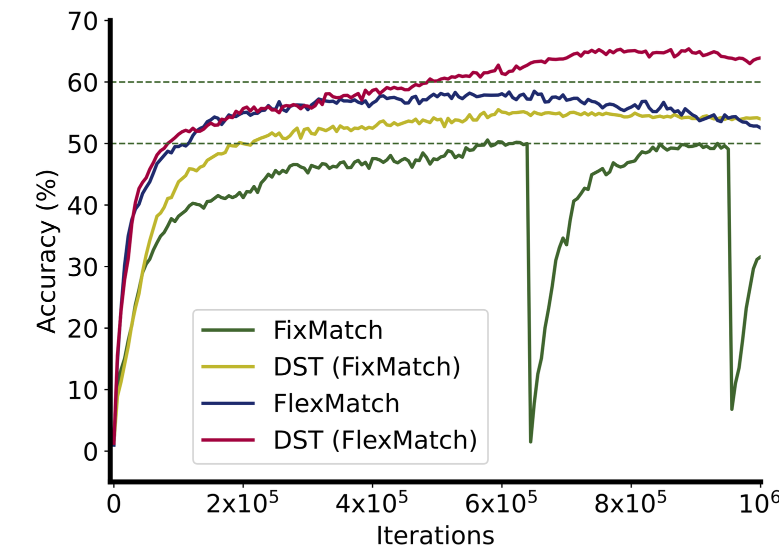
- Unify classification, self-training, and adversarial learning into a minimax game.

$$\min_{\psi,h,h_{\text{pseudo}}} \max_{h'} L_{\mathcal{L}}(\psi,h) + L_{\mathcal{U}}(\psi, h_{\text{pseudo}}, \widehat{f}_{\psi,h}) + \big(L_{\mathcal{U}}(\psi, h', \widehat{f}_{\psi,h}) - L_{\mathcal{L}}(\psi, h')\big). \quad (3)$$

Experimental Results

Standard SSL Benchmarks

Method	CIFAR-10	CIFAR-100	SVHN	STL-10	Avg
Psuedo Label	25.4	12.6	25.3	25.3	22.2
VAT	25.3	15.1	26.1	25.5	23.0
ALI	25.9	12.4	28.5	24.1	22.7
RAT	33.2	20.5	52.6	30.7	34.2
MixMatch	52.6	32.4	57.5	45.1	46.9
UDA	71.0	40.7	47.4	62.6	55.4
ReMixMatch	80.9	55.7	96.6	64.0	74.3
Dash	86.8	55.2	**97.0**	64.5	75.9
FixMatch	87.2	50.6	96.5	67.1	75.4
DST (FixMatch)	**89.3**	**56.1**	96.7	**71.0**	**78.3**
FlexMatch	94.7	59.5	89.6	71.3	78.8
DST (FlexMatch)	**95.0**	**65.4**	**94.2**	**79.6**	**83.6**

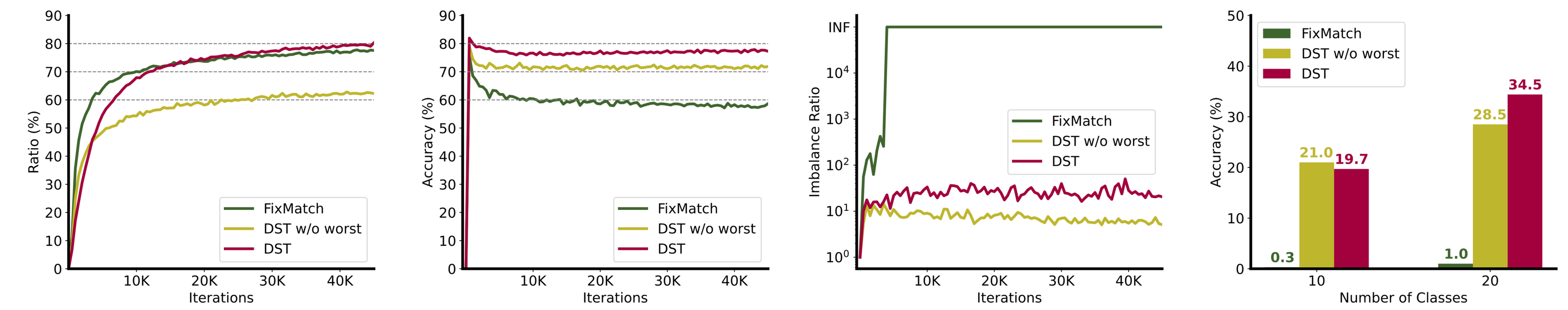


Fine-tuning from Supervised Pre-trained Models

		Caltech101	CIFAR-10	CIFAR-100	SUN397	DTD	Aircraft	CUB	Flowers	Pets	Cars	Food101	Average
Supervised	Baseline	81.4	65.2	48.2	39.9	47.7	25.4	46.5	85.2	78.1	33.3	33.8	53.2
	Pseudo Label	86.3	83.3	54.7	41.0	50.2	27.2	54.3	92.3	87.8	41.4	38.0	59.7
	Π-Model	83.5	73.1	49.2	39.7↓	50.3	24.3↓	47.1	90.7	82.2	30.9↓	33.9	55.0
	Mean Teacher	83.7	82.1	56.0	37.9↓	51.6	30.7	49.6	91.0	82.8	39.1	40.3	58.6
	VAT	84.1	72.2	48.8	39.5↓	50.6	25.9	48.1	89.4	81.8	32.4↓	36.7	55.4
	ALI	82.2	69.5	46.3↓	36.4↓	50.5	21.3↓	42.5↓	82.9↓	77.4↓	29.8↓	31.7↓	51.9
	RAT	84.0	81.8	55.4	39.0↓	49.1	31.6	50.0	89.9	84.1	37.9	38.4	58.3
	MixMatch	85.4	82.8	53.5	41.8	50.1	24.7↓	51.7	91.5	83.3	42.5	38.2	58.7
	UDA	85.8	83.6	54.7	41.3	49.0	27.1	52.1	92.0	83.1	45.6	41.7	59.6
	FixMatch	86.3	84.6	53.1	41.3	48.6	25.2↓	52.3	93.2	83.7	46.4	37.1	59.3
	Self-Tuning	87.2	76.0	57.1	41.8	50.7	35.2	58.9	92.6	86.6	58.3	41.9	62.4
	FlexMatch	87.1	89.0	63.4	48.3	52.5	34.0	54.9	94.5	88.3	57.5	49.5	65.4
	DebiasMatch	88.6	91.0	65.7	46.6	52.4	37.5	58.6	95.6	86.4	60.5	53.5	66.9
	DST (FixMatch)	89.6	94.9	70.4	48.1	53.5	43.2	68.7	94.8	89.8	71.0	**58.5**	71.1
	DST (FlexMatch)	**90.6**	**95.9**	**71.2**	**49.8**	**56.2**	**44.5**	**70.5**	**95.8**	**90.4**	**72.7**	57.1	**72.2**

- Similar results when fine-tuning from **unsupervised** pre-trained models.

How DST Improves Pseudo Labeling

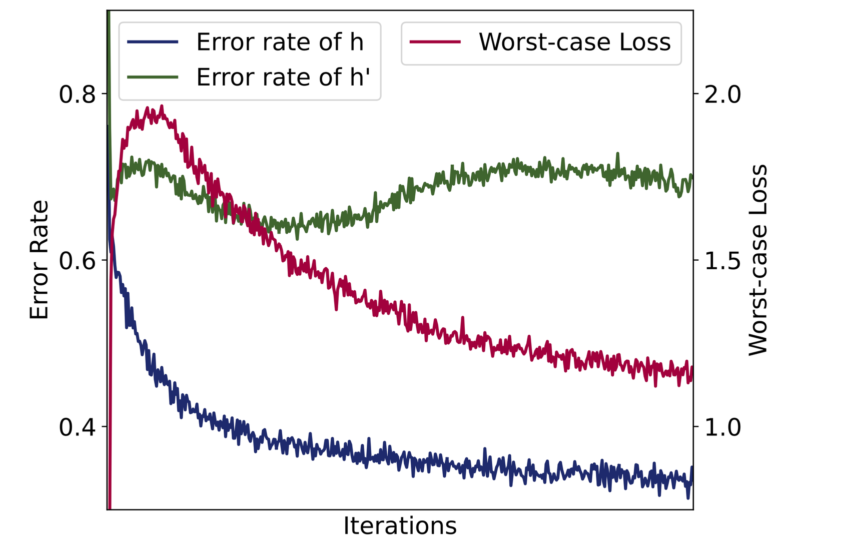


(a) Quantity (b) Quality (c) Quantity of bad classes (d) Quality of bad classes

- DST improves both the quality and quantity of pseudo labels (SubFigure (a), (b)).
- DST generates better pseudo labels for poorly-behaved classes (SubFigure (c), (d)).

DST as a General Add-on

Pre-training		Supervised		Unsupervised	
Label Amount		400	1000	400	1000
Mean Teacher	Base	56.0	67.0	51.3	63.5
	DST	**62.7**	**70.7**	**60.7**	**69.3**
Noisy Student	Base	52.8	64.3	55.6	65.8
	DST	**68.9**	**74.8**	**66.6**	**75.2**
DivideMix	Base	55.8	67.5	53.6	64.9
	DST	**69.1**	**75.1**	**65.0**	**74.2**
FixMatch	Base	53.1	67.8	51.4	64.2
	DST	**70.4**	**75.6**	**68.2**	**76.8**
FlexMatch	Base	63.4	71.2	60.2	71.1
	DST	**71.2**	**77.3**	**68.9**	**77.5**



Convergence and Computation Cost of the Minimax Game

- The worst-case error rate of h' and worst loss first increase (h' dominates), and then gradually decrease and converge (ψ dominates).
- DST introduces marginal cost (<7%) during training and **no cost during inference**.

School of Software - BNRist - Tsinghua University - China

Code: `https://github.com/thuml/Debiased-Self-Training`

Mail: `cbx22@mails.tsinghua.edu.cn`