英語論文#2

2023/06/05

M1

建元　了

論文の概要

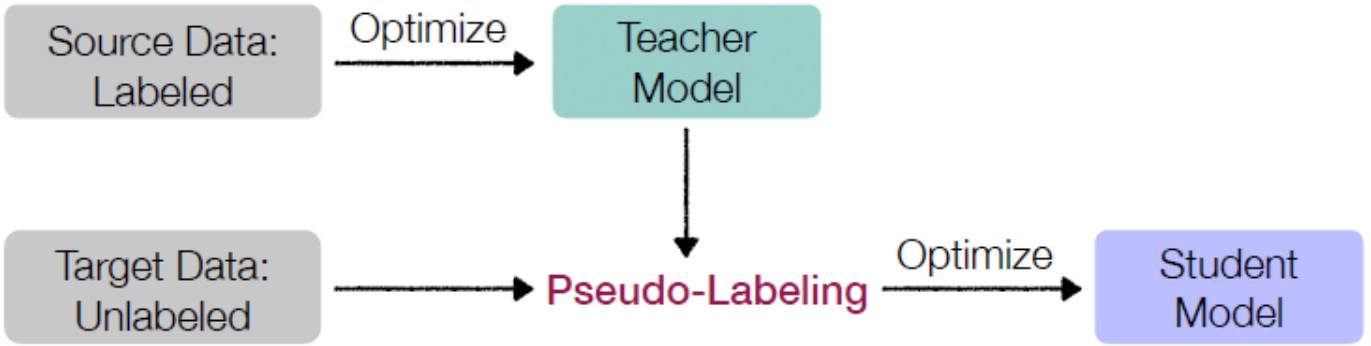
- タイトル
  - 「**Debiased Self-Training for Semi-Supervised Learnig**」

- 執筆者
  - Baixu Chen, Junguang Jiang, Ximei Wang, Pengfei Wan, Jianmin Wang, Mingsheng Long
- 掲載
  - NeurIPS 2022

- 選択理由
  - SSLにおいて予測のバイアスの調査と改善の研究であるため

背景

- **疑似ラベル**は、データセットで事前学習したモデルがラベル付かない画像に対して予測値をラベルとして扱い、データセットと組み合わせる**半教師あり学習（SSL）**の手法として広く利用

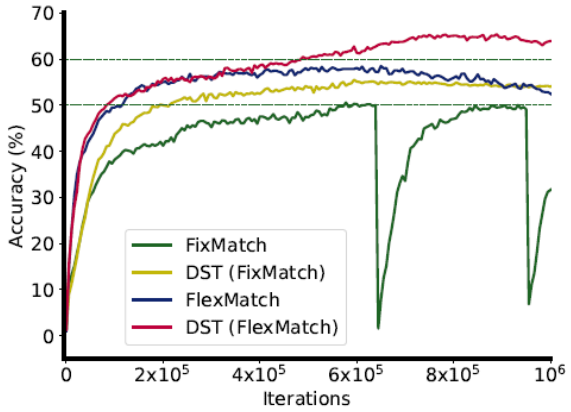


(a) Framework of pseudo-labeling

背景

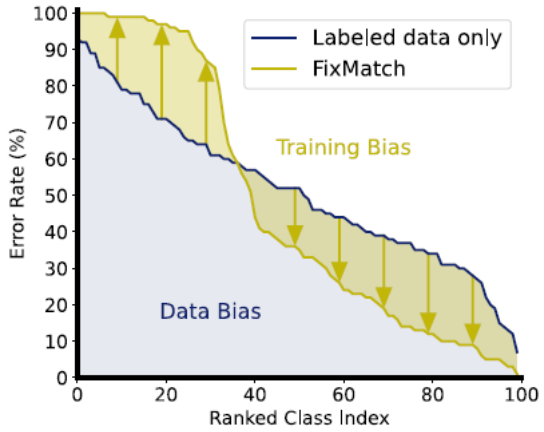
- **自己訓練（Self training）**は、依然として学習不安定性とクラス間のバイアスがある
  - Fixmatchの場合、０から学習すると大きく変動

Figure 7: Top-1 accuracy on *CIFAR-100* (train from scratch, 4 labels per category).



背景

- **Matthew effect**
  - よいデータならばさらによい精度が出て、悪いデータならばほぼゼロまで低下する
- 訓練データにクラスの不均衡が存在する場合でも、カテゴリ間の性能バランスを好む



導入

1. SSLに存在するデータバイアス
2. 誤った疑似ラベルを用いた自己訓練がもたらすバイアス増長

- **Debiased Self-Training(DST)**
  - 標準的なデータセットにおいて6.3％向上
  - 13のタスクにおいて、Fixmatchに適用したものだと18.9%の向上を達成

自己訓練におけるバイアス解析

- 入力空間Xの分布をP
- クラスKにおいてP^kは、$f(\boldsymbol{x}) = k$となる$\boldsymbol{x}$のクラス分布

- 疑似ラベルf_{pl}はn個のラベル付きサンプル$\widehat{P_n}$の学習によって得られたものとする
- 誤った疑似ラベル$M(f_{pl}) = \{\boldsymbol{x}: f_{pl}(\boldsymbol{x}) \neq f(\boldsymbol{x})\}$
- 誤った疑似ラベル付きサンプルの割合

$$B(f_{pl}) = \{P^k(M(f_{pl}))\}_{k=1}^{K}$$

自己訓練におけるバイアス解析

- **ラベル付きデータのサンプリングは自己訓練の偏りに大きく影響する**
  - データのサンプリングが異なると、同じカテゴリの精度が大きく変化
- ラベル付きデータが少ない場合、識別超平面の距離に大きな差が生じる

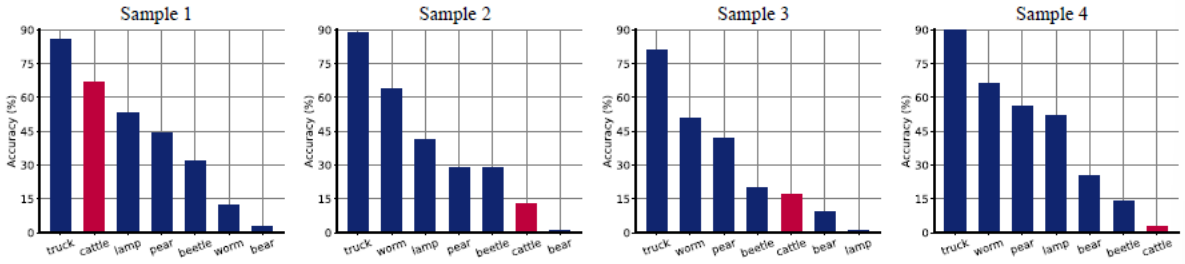


Figure 1: Effect of *data sampling*. Top-1 accuracy of 7 randomly selected categories when trained with different labeled data sampled from *CIFAR-100*. The same category (such as cattle) may have completely different accuracy in different samples. Following FixMatch [49], 4 labeled data are sampled for each category by default in our analysis.

自己訓練におけるバイアス解析

- **事前学習モデルも自己訓練バイアスに影響する**
  - 事前学習済みモデルの違いにより、カテゴリの選好性が異なる
- 同じデータでも異なる事前学習済みモデルによって、識別超平面への距離が変化することもありうる

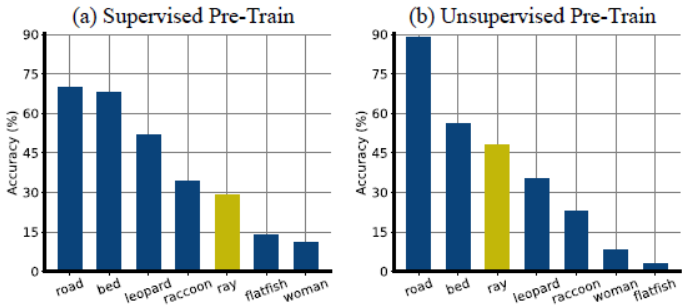


Figure 2: Effect of *pre-trained representations*. Accuracy of 7 randomly selected categories with different pre-trained models on *CIFAR-100*. Different pre-trained models show different category preferences.

自己訓練におけるバイアス解析

- **疑似ラベルを用いた学習を積極的に行うと、一部のカテゴリの自己訓練バイアスが拡大**
  - カテゴリのごとの性能差が大きくなる
- Matthew effect

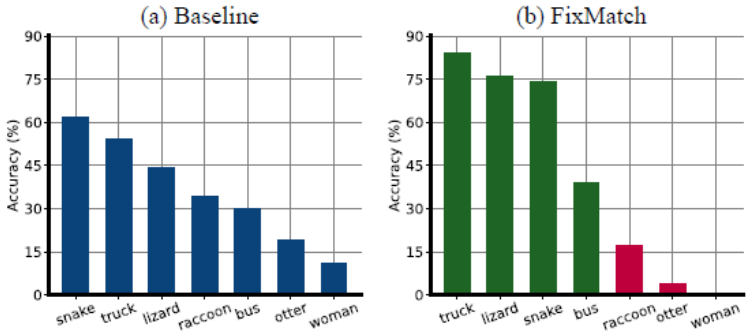


Figure 3: Effect of *self-training algorithm*. Accuracy of 7 randomly selected categories with different training methods on *CIFAR-100*. FixMatch largely increases the bias of poorly-behaved categories (Matthew effect).

自己訓練におけるバイアス解析

- データバイアス
  - SSL タスクに内在するバイアス
  - ラベルなしデータに対するサンプリングや事前学習モデルのバイアス

  - 定義

$$B\left(f_{pl}(\widehat{P_n}, \psi_0)\right)$$

  - $f_{pl}(\widehat{P_n}, \psi_0)$は偏った初期化パラメータ$\psi_0$を持つ偏ったサンプリング$\widehat{P_n}$から得る

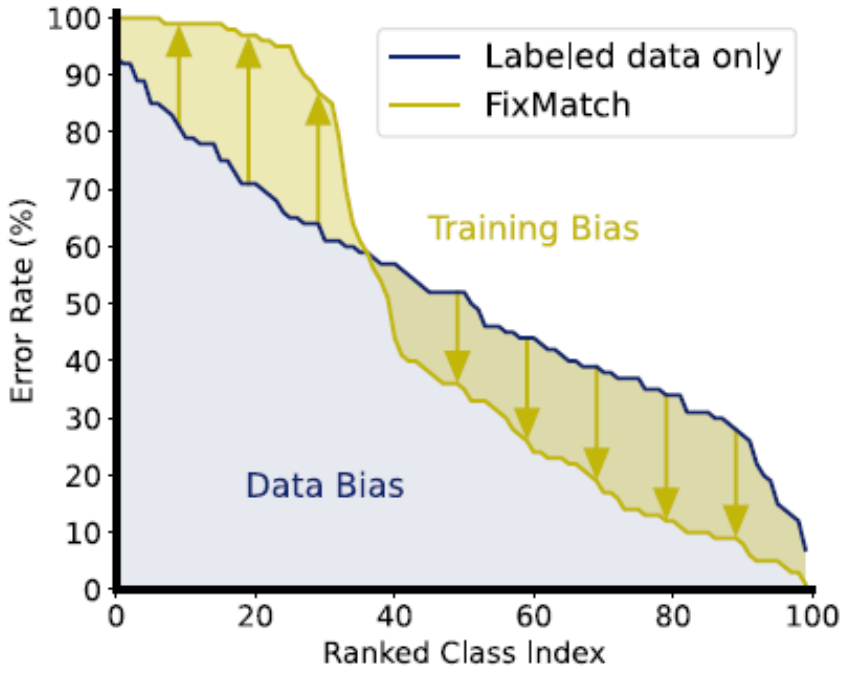
自己訓練におけるバイアス解析

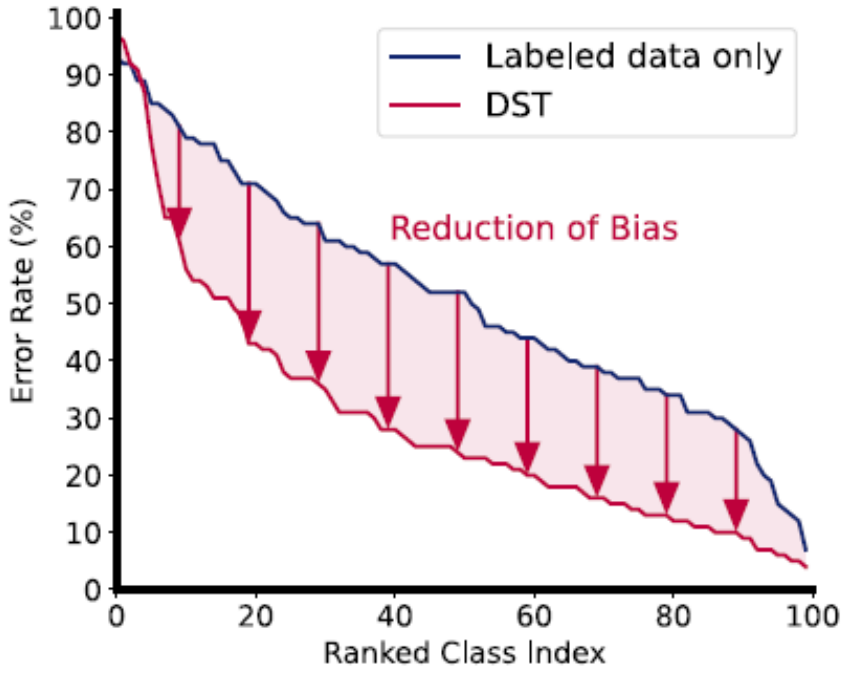
- トレーニングバイアス
  - 不合理な学習戦略によってもたらされるバイアスの拡大
  - 定義

$$B\left(f_{pl}(\widehat{P_n}, \psi_0, S)\right) - B\left(f_{pl}(\widehat{P_n}, \psi_0)\right)$$

  - $f_{pl}(\widehat{P_n}, \psi_0, S)$は自己訓練戦略$S$で得られた疑似ラベル

自己訓練におけるバイアス解析





Debiased Self-Training(DST)

- ラベル付きデータセット$\mathcal{L} = \{(x_i^l,\ y_i^l)\}_{i=1}^{n_l}$
- ラベルなしデータ$\mathcal{U} = \{(x_i^u)\}_{i=1}^{n_u},\ n_l \ll n_u$
- 特徴生成器ψ, hはタスク固有のヘッド

- 弱拡張ラベルによる交差エントロピー損失

$$L_{\mathcal{L}}(\psi, h) = \frac{1}{n_l}\sum_{i=1}^{n_l} L_{CE}((h \circ \psi \circ \alpha)(x_i^l), y_i^l)$$

Debiased Self-Training(DST)

- fixmatchは、 弱補強に対して予測値$\boldsymbol{p} = (h \circ \psi \circ \alpha)(\boldsymbol{x})$を算出、閾値$\boldsymbol{\tau}$によって疑似ラベルの信頼性を測る

$$\hat{f}_{\psi,h}(\boldsymbol{x}) = \begin{cases} \arg\max \hat{p}\,, \max \hat{p} \;\geq\; \tau \\ -1, otherwise \end{cases}$$

- $\hat{f}_{\psi,h}$は疑似ラベルを指し、これを用いて強増強のラベルなしサンプルに対して学習行う

$$L_{\mathcal{U}}\left(\psi, h, \hat{f}\right) = \frac{1}{n_l}\sum_{i=1}^{n_l} L_{CE}((h \circ \psi \circ \mathcal{A})(x_i^u), \hat{f}(x_i^u))$$

DST

- fixmatchは疑似ラベルの過程で信頼度の低いサンプルをフィルタリングするが問題がある
  1. 疑似ラベルは同じヘッドによって生成、利用されているため、トレーニングバイアスが発生
  2. 極端に少ないラベル付きサンプルで学習する場合、データの偏りに起因する信頼性の低い疑似ラベルの問題は、信頼閾値機構を用いても無視できなくなる

$$\min_{\psi,h} L_{\mathcal{L}}(\psi,h) + \lambda L_{\mathcal{U}}\left(\psi,h,\hat{f}\right)$$

➢トレーニングバイアスとデータバイアスを減少させる２つの設計の提案

DST

- より優れた教師モデルから疑似ラベルを生成し、これを利用して特徴生成器ψ、タスク固有ヘッドhの両方を学習する方法がある
  - ➢生徒モデル$h \circ \psi$の決定超平面は依然として$\hat{f}$に依存

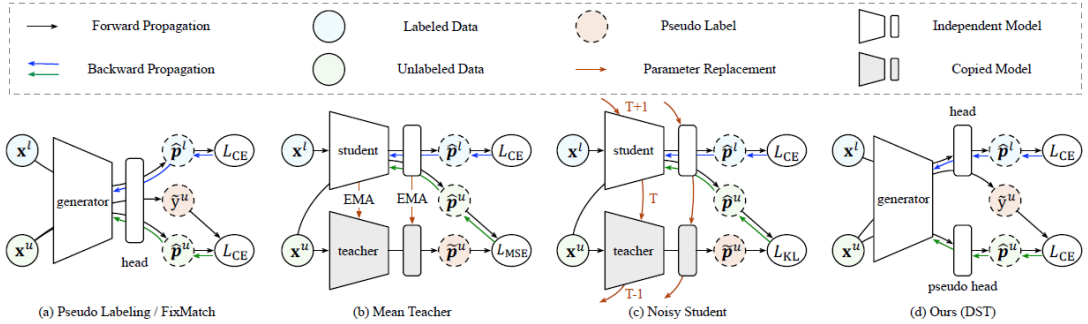
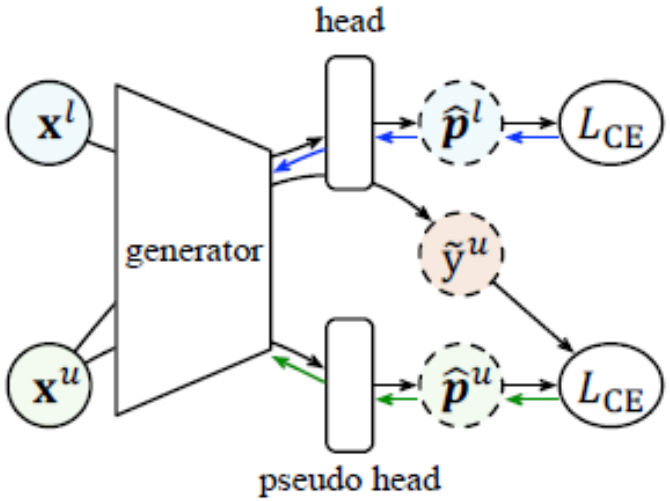


Figure 5: Comparisons on how different self-training methods generate and utilize pseudo labels. **(a)** Pseudo Labeling and FixMatch generate and utilize pseudo labels on the same model. **(b)** Mean Teacher generates pseudo labels from the Exponential Moving Average (EMA) of the current model. **(c)** Noisy Student generates pseudo labels from the teacher model which is obtained from the previous round of training. **(d)** DST generates pseudo labels from head h and utilizes pseudo labels on a parameter independent pseudo head h_{pseudo}.

DST

- トレーニングバイアスを減らす
  - ➢タスク固有ヘッドhを最適化
    - 特徴生成器ψに接続されたUから疑似ラベルのみで最適化される疑似ヘッドh_{pseudo}を導入



(d) Ours (DST)

DST

- 疑似ラベルにあるデータバイアスを減らす
  - 各クラスのラベル付きサンプルは表現空間における決定超平面への距離が異なり、ラベル付きサンプルの数が非常に少ない場合、学習した超平面と真の超平面の間に乖離が生じる
  - ➢データの偏りを減らすために特徴表現を最適化し、疑似ラベルの品質を向上させる

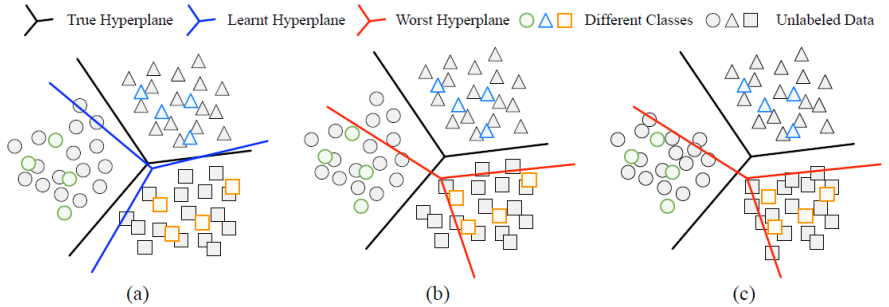


Figure 6: Concept explanations. **(a)** Shift between the hyperplanes learned on limited labeled data and the true hyperplanes. **(b)** The worst hyperplanes are hyperplanes that correctly distinguish labeled samples while making as many mistakes as possible on unlabeled samples. **(c)** Feature representations are optimized to improve the performance of the worst hyperplanes.

DST

- Uはラベルがないため、データの偏りを直接観察できない
  - トレーニングバイアスとデータバイアスの相関から見る

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

- 最終的な損失関数

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

検証

- CIFAR-10, CIFAR-100, SVHN, STL-10などSSLデータセットでランダムに初期化したDSTを評価
- 教師あり事前学習済みモデル、教師なし学習済みモデル両方を用いたDSTを使用
  1. 上位レベルの分類を含む11の下流タスクで評価
  2. 微細な分類
  3. 質感分類
  4. シーン分類

検証

- ランダムな初期化実験では WideResNetの変種を採用
- 事前学習済みモデルではImageNetで学習したResNet
- 教師なし学習事前学習済みモデルではMoCo v2を採用

実験結果

- 難易度の高いCIFAR100タスクとSTL-10タスクでは、DSTはFixMatchとFlexMatchの精度をそれぞれ8.3%と10.7%向上させた

Table 1: Top-1 accuracy on standard SSL benchmarks (train from scratch, 4 labels per category).

Method	CIFAR-10	CIFAR-100	SVHN	STL-10	Avg
Psuedo Label [30]	25.4	12.6	25.3	25.3	22.2
VAT [34]	25.3	15.1	26.1	25.5	23.0
ALI [15]	25.9	12.4	28.5	24.1	22.7
RAT [52]	33.2	20.5	52.6	30.7	34.2
MixMatch [4]	52.6	32.4	57.5	45.1	46.9
UDA [59]	71.0	40.7	47.4	62.6	55.4
ReMixMatch [3]	80.9	55.7	96.6	64.0	74.3
Dash [61]	86.8	55.2	**97.0**	64.5	75.9
FixMatch [49]	87.2	50.6	96.5	67.1	75.4
DST (FixMatch)	**89.3**	**56.1**	96.7	**71.0**	**78.3**
FlexMatch [64]	94.7	59.5	89.6	71.3	78.8
DST (FlexMatch)	**95.0**	**65.4**	**94.2**	**79.6**	**83.6**

実験結果

- FixMatchなどの典型的な自己学習法は、教師ありの事前学習モデルとの比較で比較的穏やかな改善をもたらす

- 事前学習モデルだと自己訓練の安定性に優れていた

Table 2: Comparison between DST and various baselines (ResNet50, supervised and unsupervised pre-trained, 4 labels per category). ↓ indicates a performance degradation compared with the baseline.

		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 [30]	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 [29]	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 [53]	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 [34]	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 [15]	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 [52]	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 [4]	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 [59]	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 [49]	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 [55]	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 [64]	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 [56]	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**
Unsupervised	Baseline	79.5	66.6	46.5	38.1	47.9	28.7	37.5	87.7	60.0	38.1	32.9	51.2
	Pseudo Label [30]	86.2	70.8	49.8	38.6	50.0	26.6↓	41.8	93.0	68.4	37.3↓	32.8↓	54.1
	Π-Model [29]	80.1	76.2	44.8↓	37.8↓	50.0	23.5↓	31.6↓	93.1	62.8	25.6↓	30.4↓	50.5
	Mean Teacher [53]	80.4	80.8	51.3	34.2↓	48.8	33.8	41.6	92.9	67.0	50.5	39.1	56.4
	VAT [34]	79.9	73.8	45.1↓	38.3	49.2	24.2↓	36.4↓	92.4	61.7	29.9↓	33.1	51.3
	ALI [15]	76.4↓	69.2	44.4↓	34.9↓	50.1	22.2↓	33.8↓	84.9↓	59.6↓	33.1↓	31.0↓	49.1
	RAT [52]	80.9	79.5	52.4	37.0↓	50.4	30.1	40.7	91.8	70.5	47.9	35.6	56.1
	MixMatch [4]	84.1	81.5	51.7	38.4	47.0↓	31.7	39.8	93.5	66.4	47.1	34.6	56.0
	UDA [59]	85.0	87.4	53.6	42.3	46.2↓	35.7	41.4	94.1	69.3	51.5	39.3	58.7
	FixMatch [49]	83.1	82.2	51.4	39.2	43.9↓	30.1	36.8↓	94.3	65.7	48.6	36.8	55.6
	Self-Tuning [55]	81.6	63.6↓	47.8	38.8	45.5↓	31.4	41.6	91.0	66.9	52.0	34.0	54.0
	FlexMatch [64]	86.4	96.7	60.2	45.3	53.9	42.0	49.2	95.8	72.9	69.0	37.5	64.4
	DebiasMatch [56]	86.4	96.3	66.3	44.5	53.9	44.8	51.2	95.4	70.9	72.5	53.6	66.9
	DST (FixMatch)	90.1	95.0	68.2	46.8	54.2	**47.7**	53.6	95.6	**75.4**	72.0	**57.1**	68.7
	DST (FlexMatch)	**90.4**	**96.9**	**68.9**	**48.8**	**55.9**	47.3	**55.2**	**96.4**	75.1	**74.6**	56.9	**69.7**

検証結果

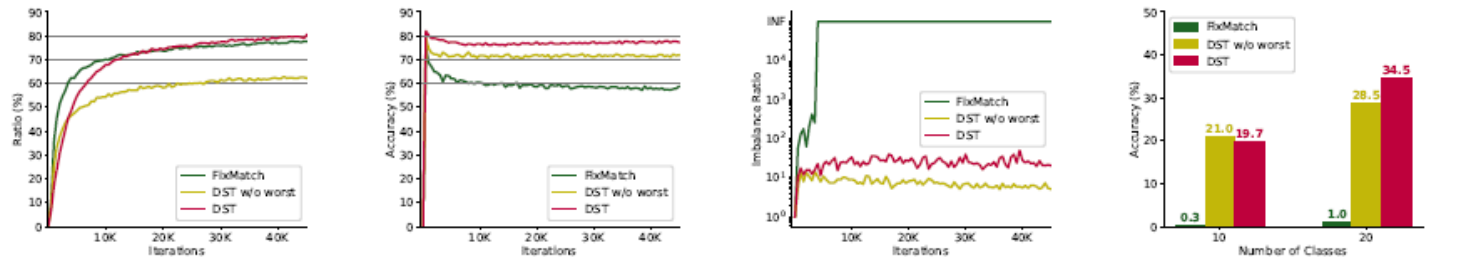
- CIFAR-100において提案手法の検証から以下の知見を得た
  1. 2つのヘッドが互いに擬似ラベルを提供し合うMutual Learningと比較し、提案手法は学習バイアスをより軽減
  2. 非線形疑似ヘッドは、線形疑似より常に優れている
  3. 疑似ラベルの推定が最悪の場合、大きなマージンによって性能向上

Table 3: Ablation study on *CIFAR-100* with different pre-trained models (4 labels per category).

Method	Multiple Heads	Linear Pseudo Head	Nonlinear Pseudo Head	Worst Case Estimation	Supervised Pre-training	Unsupervised Pre-training
FixMatch					53.1	51.4
Mutual Learning	✓				53.4	52.5
DST w/o worst	✓	✓			58.2	59.0
DST w/o worst	✓		✓		60.6	60.9
DST	✓		✓	✓	**70.4**	**68.2**

分析

- DSTは疑似ラベルの量と質の向上
  - Fixmatchは積極的にラベルなしデータを利用し、70%以上の疑似ラベルを生成するものの、誤ったクラスを多く付与する
  - DSTは量は変わらず、疑似ラベルの精度が高い



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

Figure 8: The quantity and quality of pseudo labels on *CIFAR-100* (ResNet50, supervised pre-trained).

分析

- min-max最適化と計算コスト
  - SGDを用いてψ, h'を交互に最適化
  - 4台の2080 Ti GPUを使用してCIFAR-100で1000k反復学習を行った場合、FixMatchが104時間かかるのに対し、DSTは111時間と、7%の時間増にとどまる

Figure 9: Empirical error rate and loss (*CIFAR-100*).

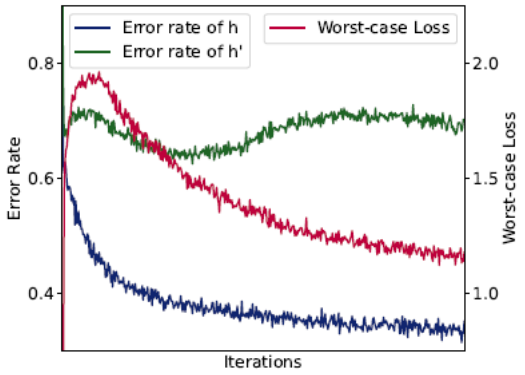


Figure 10: DST as a general add-on on *CIFAR-100*.

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**

質問補足

- SUN397
  - シーン理解（SUN）ベンチマーク
  - 397カテゴリ
  - 108753枚（訓練データ76128枚・テストデータ10875枚）


/r/rope_bridge (307)


/t/tower (358)


/b/bedroom (48)


/s/subway_station/platform (337)


/a/airport_terminal (2)


/t/toyshop (359)


/f/forest/broadleaf (159)


/a/amusement_park (6)


/c/control_tower/outdoor (110)

質問補足

- DTD（Describable Textures Dataset）
  - 人間の知覚が、テクスチャパターンから視覚的印象を引き出した後に物体の認識をすることからインスピレーションを得たデータセット
  - 47カテゴリ
  - カテゴリごとに120枚（training, validation, test）、合計5640枚