Dash: Semi-Supervised Learning with Dynamic Thresholding

Yi Xu

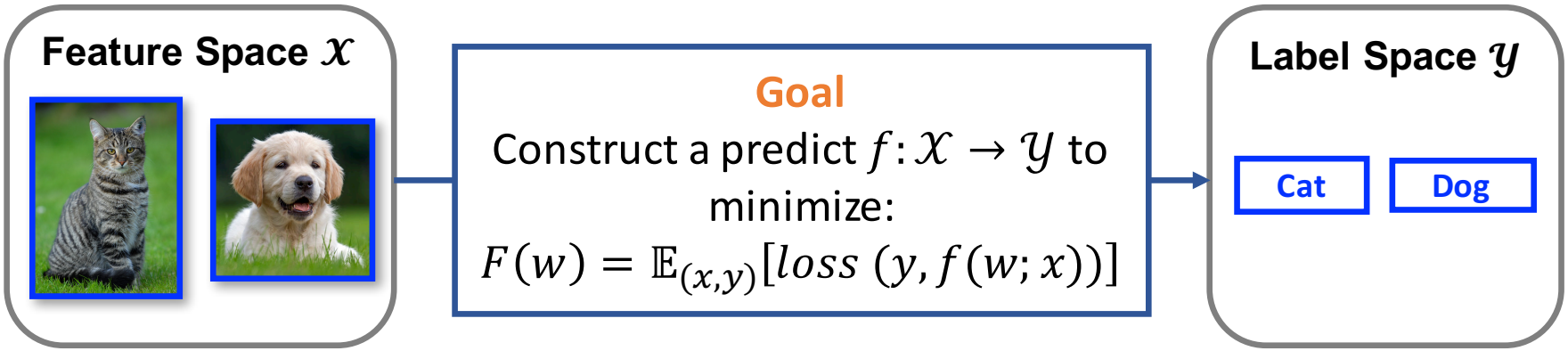
Machine Intelligence Technology, Alibaba Group

joint work with Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, Rong Jin

International Conference on Machine Learning (ICML) 2021

https://yxu71.github.io

Supervised Learning

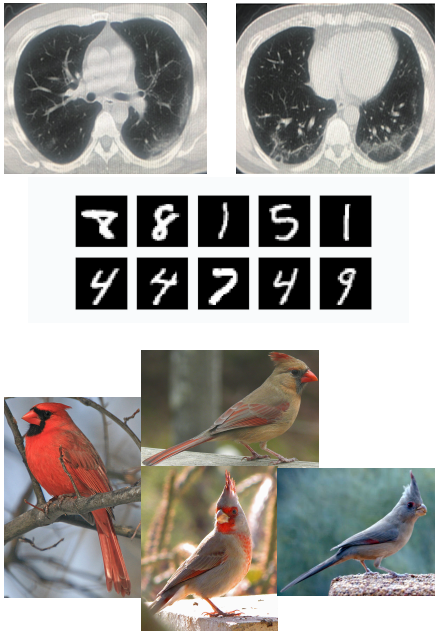


Training Data
$\{(x_i, y_i)\}_{i=1}^{n}$
Labeled

Learning Algorithm

Prediction
$\hat{f}$

Labeled and Unlabeled Data



Unlabeled Data

cheap & abundant

Human expert
Special equipment
Experiment

“Covid-19 positive”
“Covid-19 negative”

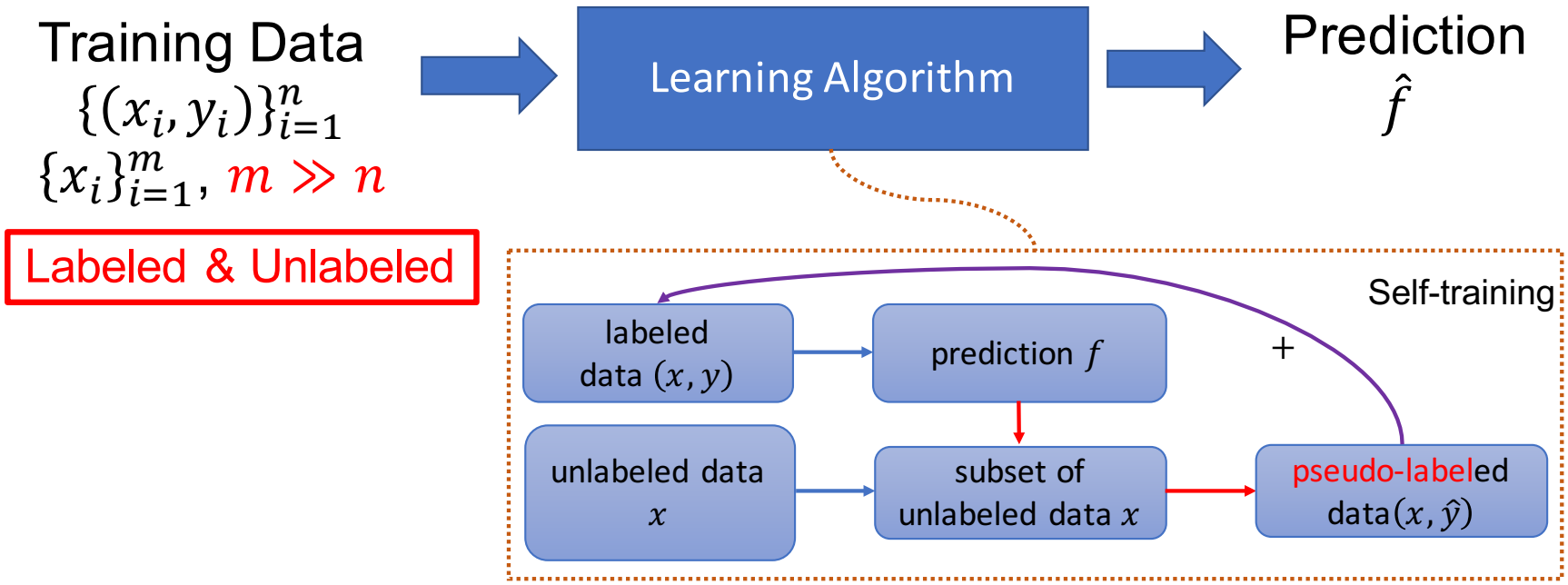
“8”, “8”, “1”, “5”, “1”
“4”, “4”, “7”, “4”, “9”

“Northern cardinal (M)”
“Northern cardinal (F)”
“desert cardinal (M)”
“desert cardinal (F)”

Labeled Data

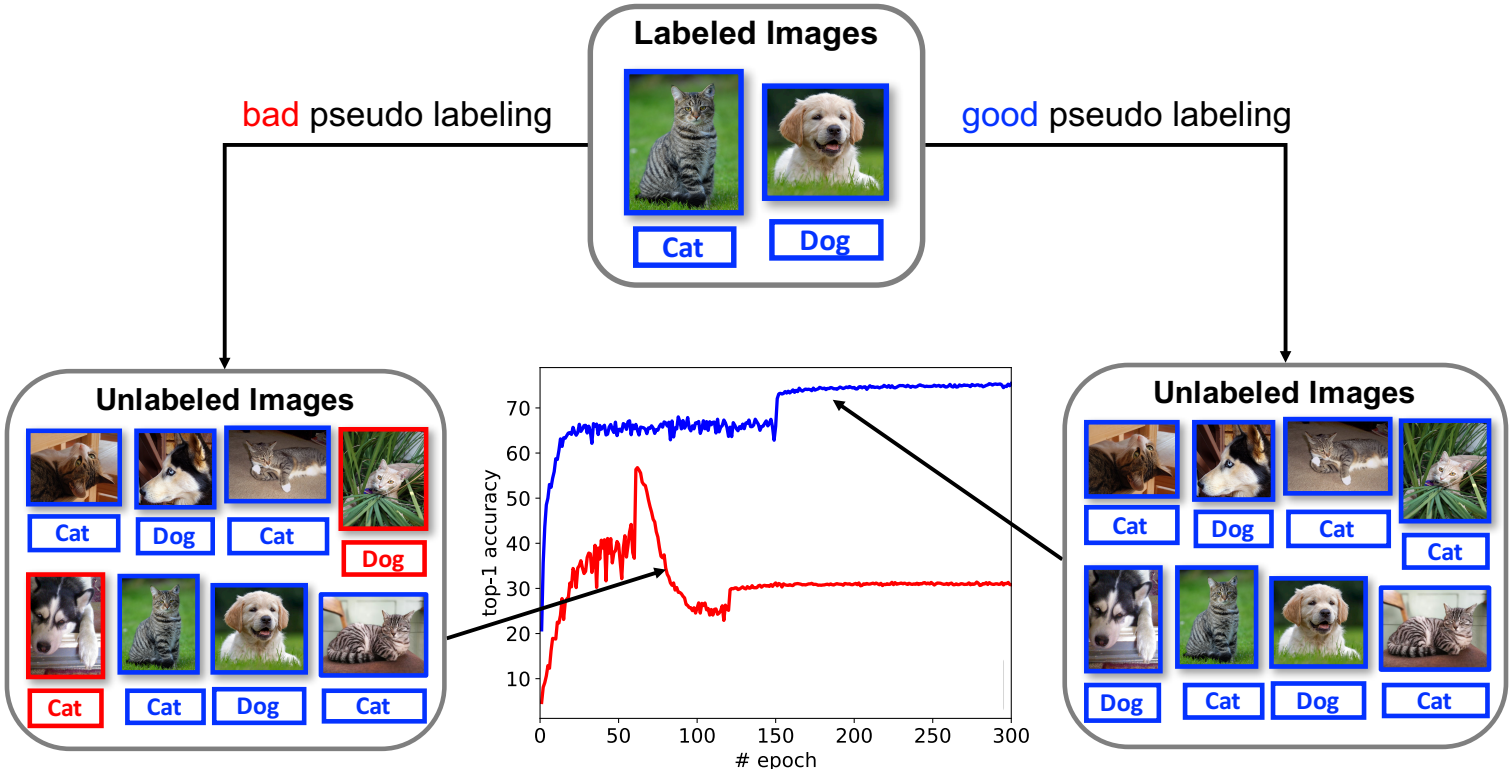
expensive & scarce

Semi-Supervised Learning (SSL)



Goal: Learn a better prediction than based on labeled data alone

An Issue of Unlabeled Data



Not All Unlabled Data are Needed!

Exisiting SSL methods:

- use all unlabeled examples
- use the unlabeled examples with a fixed high-confidence prediction

FixMatch

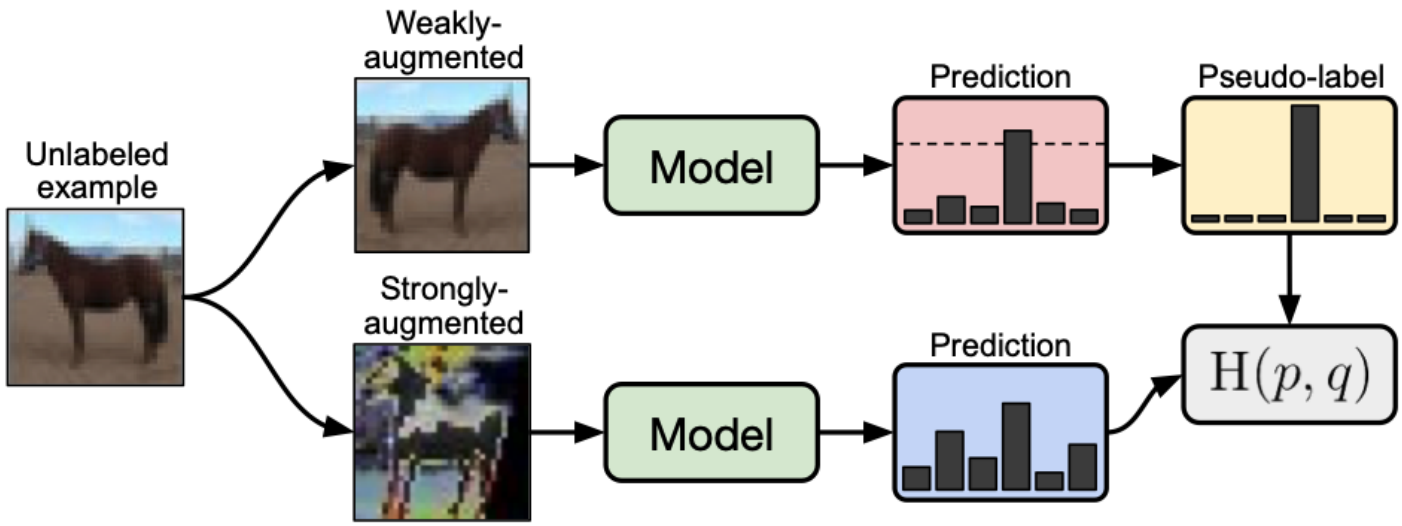
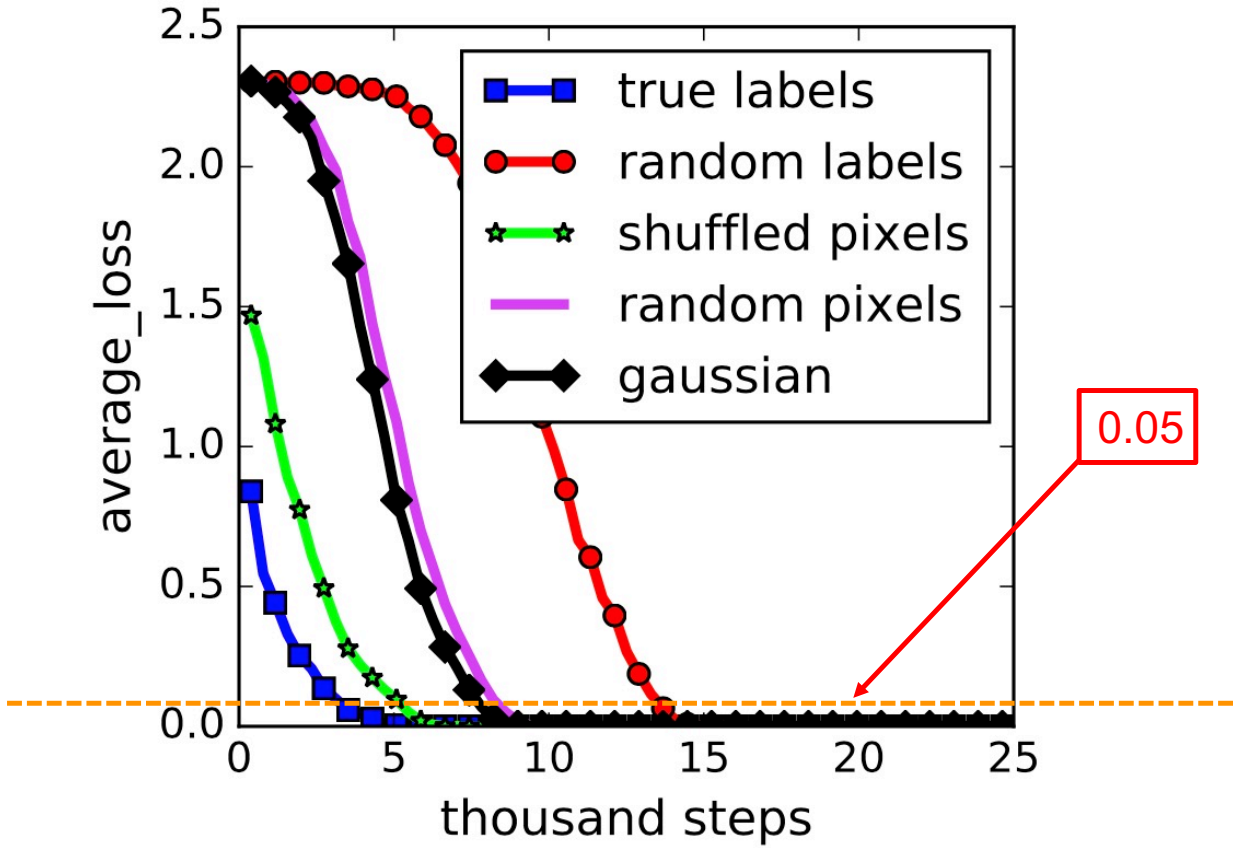


Figure: Diagram of FixMatch (Sohn et al., 2020)

- Weak augmentation: standard flip-and-shift augmentation
- Strong augmentation: RandAugment (RA) (Cubuk et al., 2020) or CTAugment (CTA) (Cubuk et al., 2019)
- FixMatch uses the unlabeled examples with a fixed high-confidence prediction, i.e., the probability ≥ 0.95.
  - one-hot loss $\leq -\log(0.95) \approx 0.05$

Fixed threshold is not good enough

The curve of training loss in deep learning (Zhang et al., 2016):

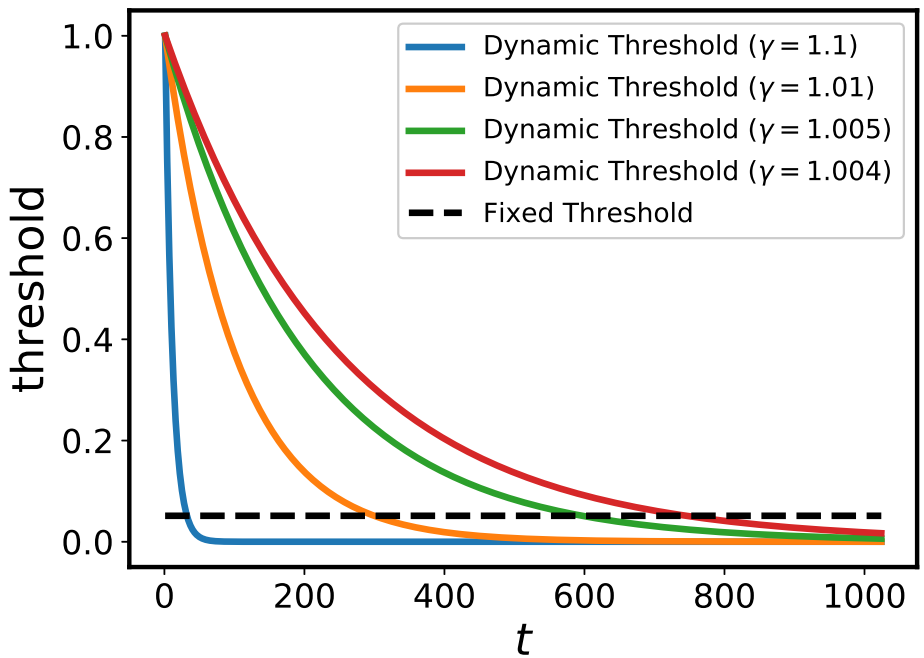


Our **Dash** method: Dynamic threshold

The dynamic threshold ρ_t is a decreasing function of t:

$$\rho_t := C\gamma^{-(t-1)}\widehat{\rho}, \quad C > 1, \gamma > 1. \tag{1}$$

- In all experiments, $C = 1.0001$.
- Tuning parameter γ. Specifically, $\gamma = 1.27$ in our experiments.
- $\widehat{\rho} \approx$ training loss on labeled data after first epoch



Dynamic threshold vs. Fixed threshold

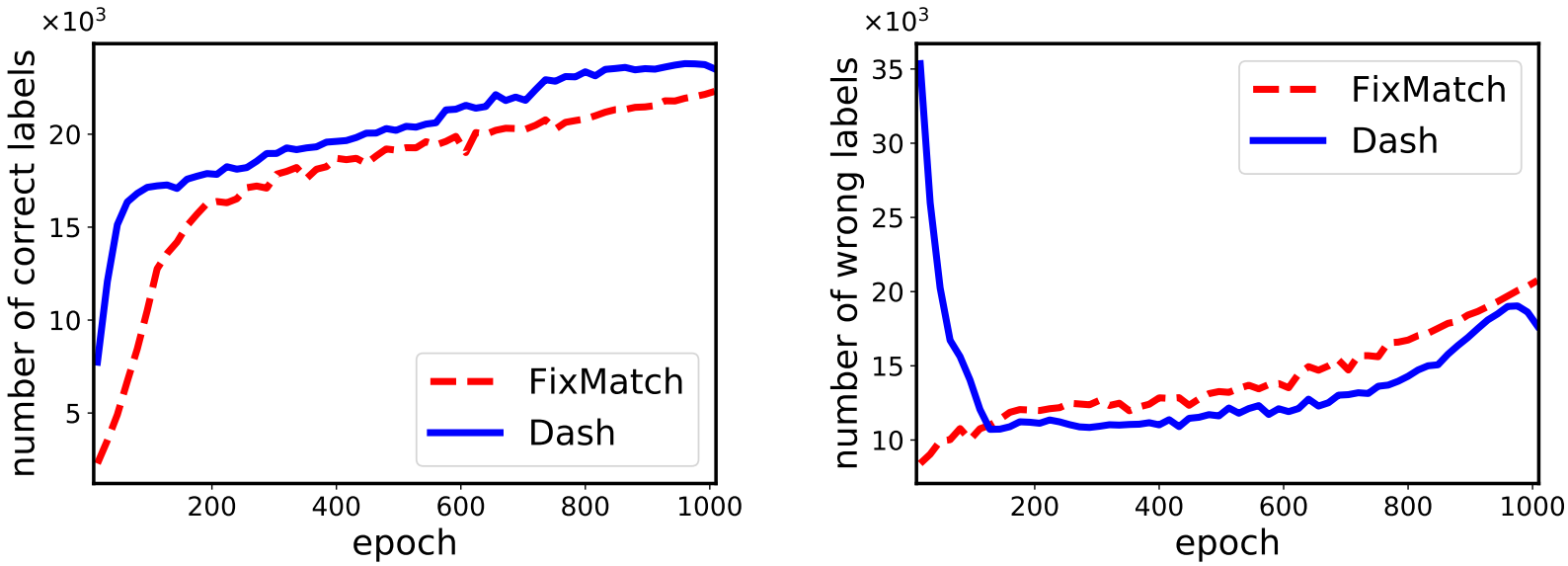


Figure: Number of selected unlabeled images with correct/wrong pseudo labels

The proposed Dash algorithm

1: // Warm-up Stage: run SGD in T_0 iterations.
2: Initialization: $\mathbf{u}_0 = \mathbf{w}_0$
3: **for** $t = 0, 1, \ldots, T_0 - 1$ **do**
4: $\quad \mathbf{u}_{t+1} = \mathbf{u}_t - \eta_0 \frac{1}{m_0} \sum_{i=1}^{m_0} \nabla f(\mathbf{u}_t; \xi_{t,i})$
5: **end for**
6: // Selection Stage: run SGD in T iterations.
7: Initialization: $\mathbf{w}_1 = \mathbf{u}_{T_0}$.
8: $\widehat{\rho} = \frac{1}{n} \sum_{\xi_i \in \mathcal{D}_l} f(\mathbf{w}_1; \xi_i)$
9: **for** $t = 1, \ldots, T$ **do**
10: $\quad$ 1) Sample m examples from $\mathcal{D}_u$ (pseudo labels in $\mathcal{D}_u$ are generated by FixMatch)
11: $\quad$ 2) $\mathbf{g}_t = \frac{\sum_{i=1}^{m} I(f_u(\mathbf{w}_t; \xi_{t,i}^u) \leq \rho_t) \nabla f_u(\mathbf{w}_t; \xi_{t,i}^u)}{\sum_{i=1}^{m} I(f_u(\mathbf{w}_t; \xi_{t,i}^u) \leq \rho_t)}$, where $\rho_t = C\gamma^{-(t-1)}\widehat{\rho}$
12: $\quad$ 3) $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{g}_t$.
13: **end for**
14: Output: $\mathbf{w}_{T+1}$

Convergence result

- Non-convex problem: $\min_{\mathbf{w}\in\mathbb{R}} F(\mathbf{w}) = \mathrm{E}_{\xi}[f(\mathbf{w};\xi)]$
- Optimization method: SGD
- Iteration complexity: number of iterations $T(\epsilon)$ that $F(\mathbf{w}_T) - F(\mathbf{w}_*) \leq \epsilon$, where $0 < \epsilon \ll 1$, (e.g. 10^{-3}).
- Supervised learning: optimal complexity is $T(\epsilon) = O\left(\frac{1}{\epsilon^4}\right)$
- Semi-supervised learning: unclear
  - Under "perfect" pseudo-labeling, same complexity as supervised learning
  - Challenge: e.g., out-of-distribution, wrong pseudo labels
- Under some assumptions, our convergence result for Dash:

$$T(\epsilon) = O\left(\frac{1}{\epsilon^4}\right)$$

Experiments

Comparison of top-1 testing error rates

	SVHN			STL-10
Algorithm	40 labels	250 labels	1000 labels	1000 labels
Π-Model	-	18.96±1.92	7.54±0.36	26.23±0.82
Pseudo-Labeling	-	20.21±1.09	9.94±0.61	27.99±0.83
Mean Teacher	-	3.57±0.11	3.42±0.07	21.43±2.39
MixMatch	42.55±14.53	3.98±0.23	3.50±0.28	10.41±0.61
UDA	52.63±20.51	5.69±2.76	2.46±0.24	7.66±0.56
ReMixMatch	3.34±0.20	2.92±0.48	2.65±0.08	5.23±0.45
RYS (UDA)	-	2.45±0.08	2.32±0.06	-
RYS (FixMatch)	-	2.63±0.23	2.34±0.15	-
FixMatch (CTA)	7.65±7.65	2.64±0.64	2.36±0.19	5.17±0.63
Dash (CTA, ours)	3.14±1.60	2.38±0.29	2.14±0.09	**3.96**±0.25
FixMatch (RA)	3.96±2.17	2.48±0.38	2.28±0.11	7.98±1.50
Dash (RA, ours)	**3.03**±1.59	**2.17**±0.10	**2.03**±0.06	7.26±0.40

	CIFAR-10			CIFAR-100		
Algorithm	40 labels	250 labels	4000 labels	400 labels	2500 labels	10000 labels
Π-Model	-	54.26±3.97	14.01±0.38	-	57.25±0.48	37.88±0.11
Pseudo-Labeling	-	49.78±0.43	16.09±0.28	-	57.38±0.46	36.21±0.19
Mean Teacher	-	32.32±2.30	9.19±0.19	-	53.91±0.57	35.83±0.24
MixMatch	47.54±11.50	11.05±0.86	6.42±0.10	67.61±1.32	39.94±0.37	28.31±0.33
UDA	29.05±5.93	8.82±1.08	4.88±0.18	59.28±0.88	33.13± 0.22	24.50±0.25
ReMixMatch	19.10±9.64	5.44±0.05	4.72±0.13	**44.28**±2.06	27.43±0.31	23.03±0.56
RYS (UDA)	-	5.53±0.17	4.75±0.28	-	-	-
RYS (FixMatch)	-	5.05±0.12	4.35±0.06	-	-	-
FixMatch (CTA)	11.39±3.35	5.07±0.33	4.31±0.15	49.95±3.01	28.64±0.24	23.18±0.11
Dash (CTA, ours)	**9.16**±4.31	4.78±0.12	4.13±0.06	44.83±1.36	27.85±0.19	22.77±0.21
FixMatch (RA)	13.81±3.37	5.07±0.65	4.26±0.05	48.85±1.75	28.29±0.11	22.60±0.12
Dash (RA, ours)	13.22±3.75	**4.56**±0.13	**4.08**±0.06	44.76±0.96	**27.18**±0.21	**21.97**±0.14

Experiments

Comparison of top-1 testing error rates

	SVHN			STL-10
Algorithm	40 labels	250 labels	1000 labels	1000 labels
Π-Model	-	18.96±1.92	7.54±0.36	26.23±0.82
Pseudo-Labeling	-	20.21±1.09	9.94±0.61	27.99±0.83
Mean Teacher	-	3.57±0.11	3.42±0.07	21.43±2.39
MixMatch	42.55±14.53	3.98±0.23	3.50±0.28	10.41±0.61
UDA	52.63±20.51	5.69±2.76	2.46±0.24	7.66±0.56
ReMixMatch	3.34±0.20	2.92±0.48	2.65±0.08	5.23±0.45
RYS (UDA)	-	2.45±0.08	2.32±0.06	-
RYS (FixMatch)	-	2.63±0.23	2.34±0.15	-
FixMatch (CTA)	7.65±7.65	2.64±0.64	2.36±0.19	5.17±0.63
Dash (CTA, ours)	3.14±1.60	2.38±0.29	2.14±0.09	**3.96**±0.25
FixMatch (RA)	3.96±2.17	2.48±0.38	2.28±0.11	7.98±1.50
Dash (RA, ours)	**3.03**±1.59	**2.17**±0.10	**2.03**±0.06	7.26±0.40

	CIFAR-10			CIFAR-100		
Algorithm	40 labels	250 labels				els
Π-Model	-	54.26±3.97				11
Pseudo-Labeling	-	49.78±0.43	16.09±0.28	-	57.38±0.46	36.21±0.19
Mean Teacher	-	32.32±2.30	9.19±0.19		53.91±0.57	35.83±0.24
MixMatch	47.54±11.50	11.05±0.86	6.42±0.10	67.61±1.32	39.94±0.37	28.31±0.33
UDA	29.05±5.93	8.82±1.08	4.88±0.18	59.28±0.88	33.13± 0.22	24.50±0.25
ReMixMatch	19.10±9.64	5.44±0.05	4.72±0.13	**44.28**±2.06	27.43±0.31	23.03±0.56
RYS (UDA)	-	5.53±0.17	4.75±0.28	-	-	-
RYS (FixMatch)	-	5.05±0.12	4.35±0.06	-	-	-
FixMatch (CTA)	11.39±3.35	5.07±0.33	4.31±0.15	49.95±3.01	28.64±0.24	23.18±0.11
Dash (CTA, ours)	**9.16**±4.31	4.78±0.12	4.13±0.06	44.83±1.36	27.85±0.19	22.77±0.21
FixMatch (RA)	13.81±3.37	5.07±0.65	4.26±0.05	48.85±1.75	28.29±0.11	22.60±0.12
Dash (RA, ours)	13.22±3.75	**4.56**±0.13	**4.08**±0.06	44.76±0.96	**27.18**±0.21	**21.97**±0.14

Dash with distribution alignment: **43.31**

Conclusion

- Propose a SSL algorithm with dynamic thresholding Dash.
- Establish the convergence guarantee for Dash.
- Experimental results evaluate the efficacy of Dash.

References

Cubuk, Ekin D, Zoph, Barret, Mane, Dandelion, Vasudevan, Vijay, and Le, Quoc V. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 113–123, 2019.

Cubuk, Ekin D, Zoph, Barret, Shlens, Jonathon, and Le, Quoc V. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020.

Sohn, Kihyuk, Berthelot, David, Li, Chun-Liang, Zhang, Zizhao, Carlini, Nicholas, Cubuk, Ekin D, Kurakin, Alex, Zhang, Han, and Raffel, Colin. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

Zhang, Chiyuan, Bengio, Samy, Hardt, Moritz, Recht, Benjamin, and Vinyals, Oriol. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.