Self-Tuning for Data-Efficient Deep Learning

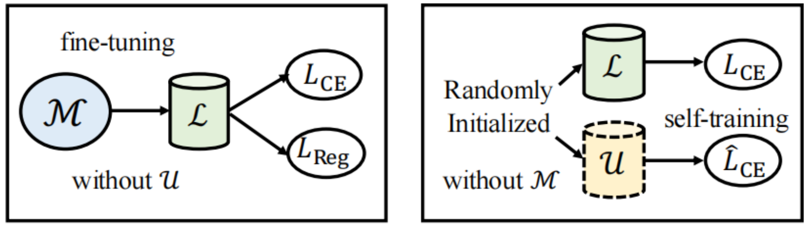
Ximei Wang [*1] **Jinghan Gao** [*1] **Mingsheng Long** [1] **Jianmin Wang** [1]

School of Software, BNRist, Tsinghua University, Beijing, China, 100084

ICML 2021

SSL & TL

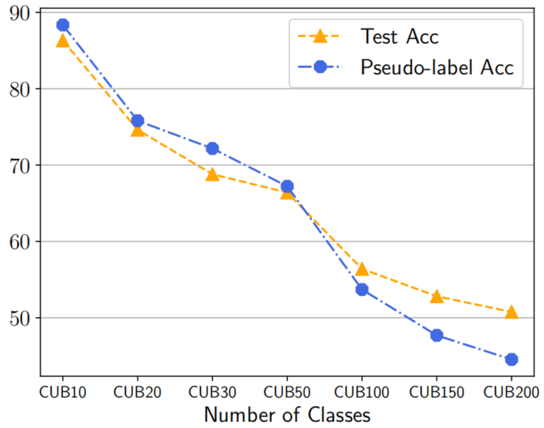
- SSL (pseudo labels) and Confirmation bias
  - Without a decent pre-trained model to provide an implicit regularization, will be easily misled by inaccurate pseudo-labels, especially in large-sized label space.
  - SSL need a well pre-trained model
- TL and Model shift
  - The fine-tuned model shifts towards the limited labeled data and leaves away from the original smooth model pre-trained on a large-scale datasets
  - Utilizing unlabeled data to alleviate the model shifting



(a) **Transfer Learning** (b) **Semi-supervised Learning**

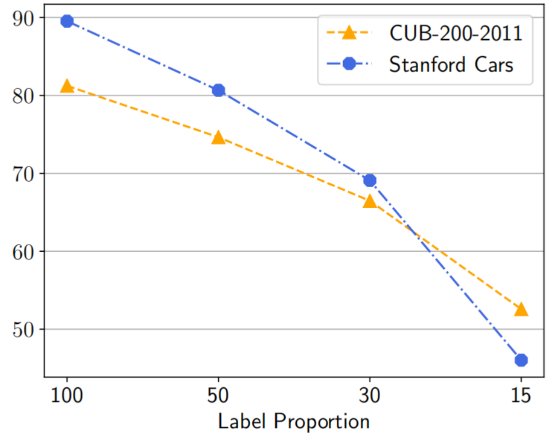
SSL and TL are complementary

SSL



(a) Acc of FixMatch on *CUB*

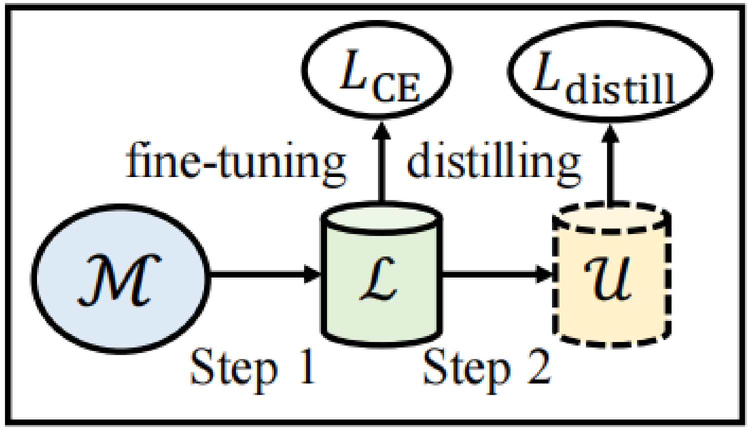
TL



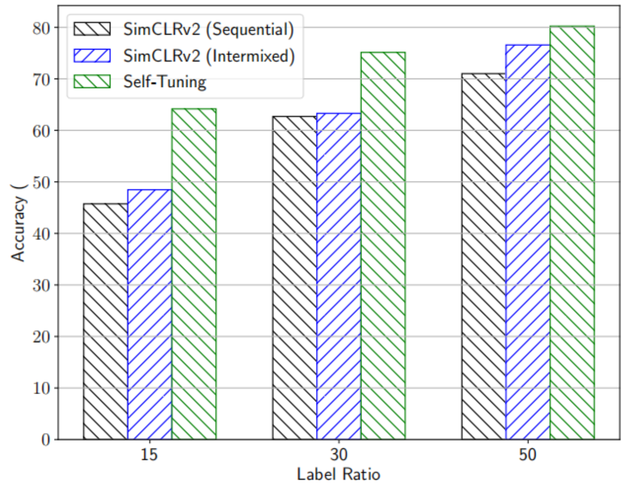
(b) Test accuracy of Co-Tuning

SimCLRv2

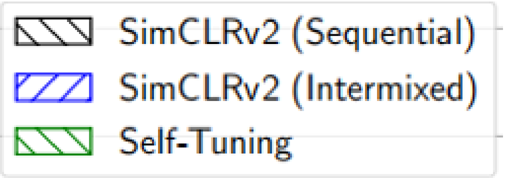
- Both utilize labeled data and unlabeled data
- The fine-tuned model would easily shift towards the limited labeled data with sampling bias and leaves away from the original smooth model pre-trained on a large-scale dataset



(c) **SimCLRv2**



(b) Compare with SimCLRv2



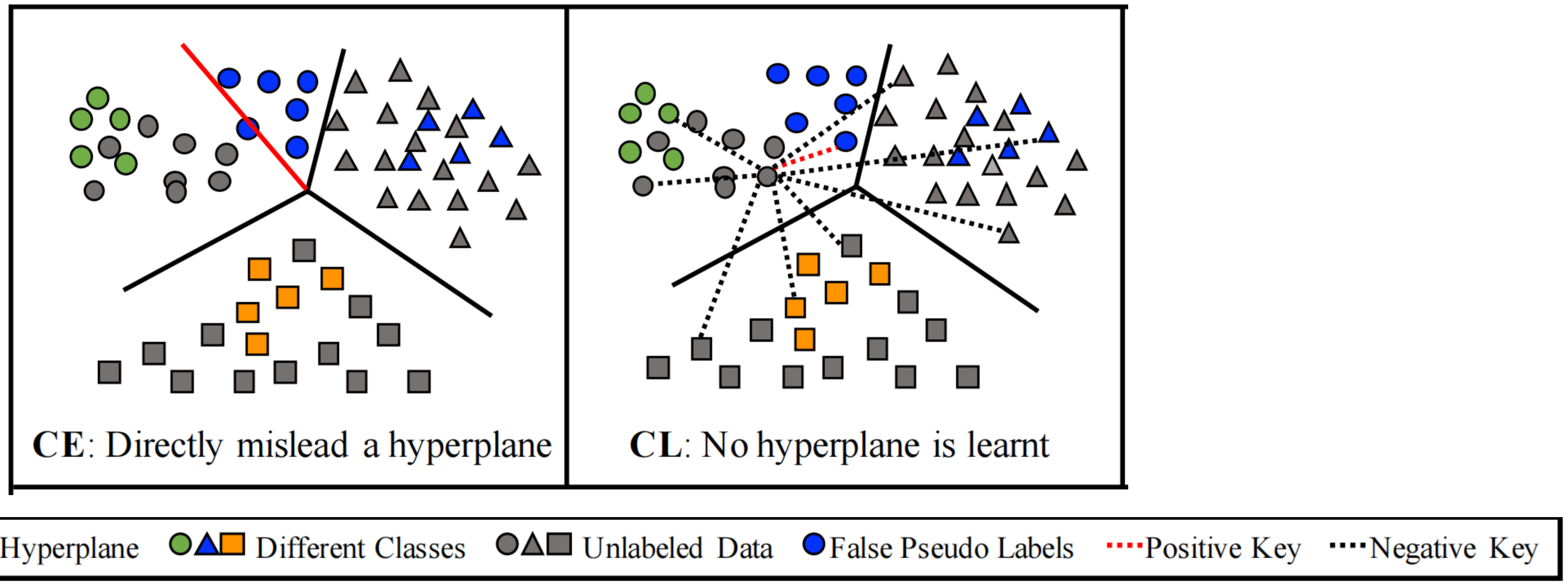
Confirmation bias: CE & CL loss

- **Cross Entropy**

  The model trained by CE loss will be easily confused by false pseudo-labels since it focuses on learning a hyperplane for discriminating each class from the other classes (CE loss overfitting easily)

- **Contrast Loss**

  While standard CL loss lacks a mechanism to tailor pseudo-labels into model training, leaving the useful discriminative information on the shelf. (CL doesn’t take classes into account)



Confirmation bias Solution-PGC

- Different from the standard CL which involves just a positive key in each contrast, PGC introduces a group of positive keys in the same pseudo-class to contrast with all negative keys from other pseudo-classes.

Queue size

Two different augments of one sample

$$\widehat{L}_{\mathrm{PGC}} = -\frac{1}{D+1}\sum_{d=0}^{\boxed{D}}\log\frac{\exp(\boxed{\mathbf{q}\cdot\mathbf{k}_d^{\widehat{y}}}/\tau)}{\mathrm{Pos}+\mathrm{Neg}}$$

Unlabeled Data

$$\mathrm{Pos} = \exp(\mathbf{q}\cdot\boxed{\mathbf{k}_0^{\widehat{y}}}/\tau) + \sum_{j=1}^{D}\exp(\mathbf{q}\cdot\mathbf{k}_j^{\widehat{y}}/\tau)$$

The second augments of queried sample

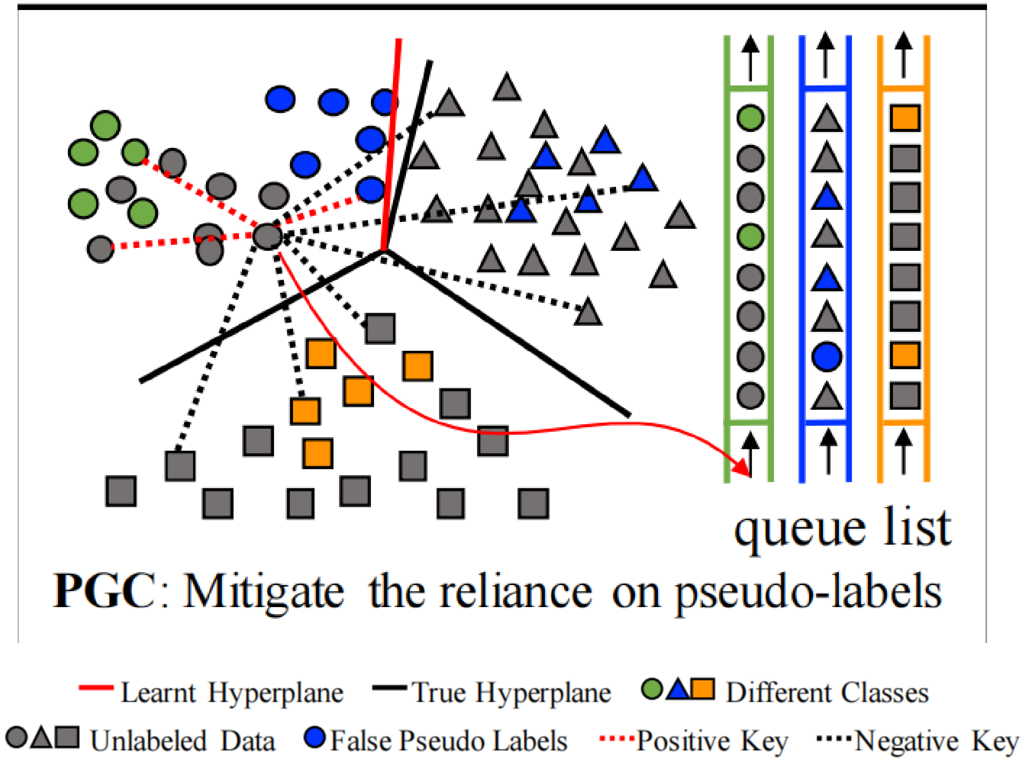
$$\mathrm{Neg} = \sum_{c=1}^{\{1,2,\cdots,C\}\setminus\widehat{y}}\sum_{j=1}^{D}\exp(\mathbf{q}\cdot\mathbf{k}_j^{c}/\tau),$$

Labeled Data

$$L_{\mathrm{PGC}} = -\frac{1}{D+1}\sum_{d=0}^{D}\log\frac{\exp(\mathbf{q}\cdot\mathbf{k}_d^{y}/\tau)}{\mathrm{Pos}+\mathrm{Neg}},$$

Loss

$$\mathbb{E}_{(\mathbf{x}_i,y_i)\in\mathcal{L}}\left(L_{\mathrm{CE}}+L_{\mathrm{PGC}}\right)+\mathbb{E}_{(\mathbf{x}_i)\in\mathcal{U}}\widehat{L}_{\mathrm{PGC}}.$$

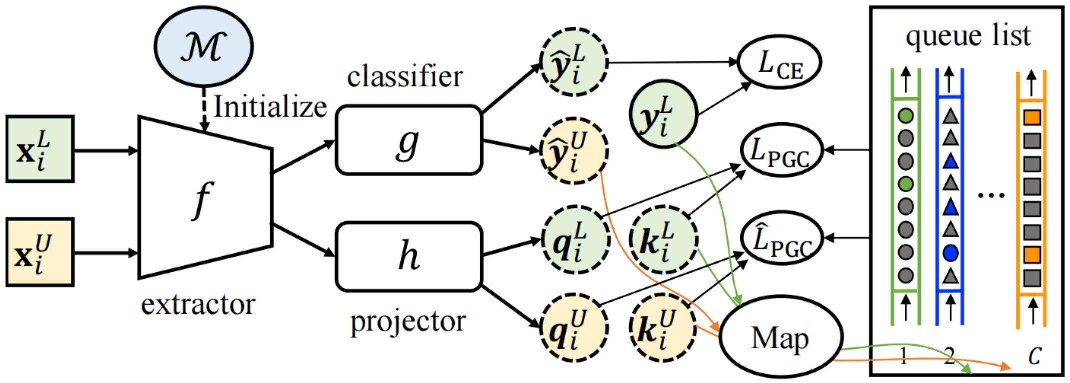


Model Shift Solution-Unifying and Sharing

- By utilizing Unlabeled data at the same time in a unified form as shown the model shift challenge is expected to be alleviated.

- Shared queue improves the accuracy keys for unlabeled queries than that of a separate queue for unlabeled data.

- Self-Tuning has a better starting point to provide an implicit regularization than the model trained from scratch on the target dataset.



Experiments

Dataset	Type	Method	Label Proportion			
			15%	30%	50%	100%
CUB-200-2011	TL	Fine-Tuning (baseline)	45.25±0.12	59.68±0.21	70.12±0.29	78.01±0.16
		L^2-SP (Li et al., 2018)	45.08±0.19	57.78±0.24	69.47±0.29	78.44±0.17
		DELTA (Li et al., 2019)	46.83±0.21	60.37±0.25	71.38±0.20	78.63±0.18
		BSS (Chen et al., 2019)	47.74±0.23	63.38±0.29	72.56±0.17	78.85±0.31
		Co-Tuning (You et al., 2020)	52.58±0.53	66.47±0.17	74.64±0.36	81.24±0.14
	SSL	Π-model (Laine & Aila, 2017)	45.20±0.23	56.20±0.29	64.07±0.32	–
		Pseudo-Labeling (Lee, 2013)	45.33±0.24	62.02±0.31	72.30±0.29	–
		Mean Teacher (Tarvainen & Valpola, 2017)	53.26±0.19	66.66±0.20	74.37±0.30	–
		UDA (Xie et al., 2020)	46.90±0.31	61.16±0.35	71.86±0.43	–
		FixMatch (Sohn et al., 2020)	44.06±0.23	63.54±0.18	75.96±0.29	–
		SimCLRv2 (Chen et al., 2020b)	45.74±0.15	62.70±0.24	71.01±0.34	–
	Combine	Co-Tuning + Pseudo-Labeling	54.11±0.24	68.07±0.32	75.94±0.34	–
		Co-Tuning + Mean Teacher	57.92±0.18	67.98±0.25	72.82±0.29	–
		Co-Tuning + FixMatch	46.81±0.21	58.88±0.23	73.07±0.29	–
		Self-Tuning (ours)	**64.17**±0.47	**75.13**±0.35	**80.22**±0.36	**83.95**±0.18

Experiments

Dataset	Type	Method	Label Proportion			
			15%	30%	50%	100%
Stanford Cars	TL	Fine-Tuning (baseline)	36.77±0.12	60.63±0.18	75.10±0.21	87.20±0.19
		L^2-SP (Li et al., 2018)	36.10±0.30	60.30±0.28	75.48±0.22	86.58±0.26
		DELTA (Li et al., 2019)	39.37±0.34	63.28±0.27	76.53±0.24	86.32±0.20
		BSS (Chen et al., 2019)	40.57±0.12	64.13±0.18	76.78±0.21	87.63±0.27
		Co-Tuning (You et al., 2020)	46.02±0.18	69.09±0.10	80.66±0.25	89.53±0.09
	SSL	Π-model (Laine & Aila, 2017)	45.19±0.21	57.29±0.26	64.18±0.29	–
		Pseudo-Labeling (Lee, 2013)	40.93±0.23	67.02±0.19	78.71±0.30	–
		Mean Teacher (Tarvainen & Valpola, 2017)	54.28±0.14	66.02±0.21	74.24±0.23	–
		UDA (Xie et al., 2020)	39.90±0.43	64.16±0.40	71.86±0.56	–
		FixMatch (Sohn et al., 2020)	49.86±0.27	77.54±0.29	84.78±0.33	–
		SimCLRv2 (Chen et al., 2020b)	45.74±0.16	61.70±0.18	77.49±0.24	–
	Combine	Co-Tuning + Pseudo-Labeling	50.16±0.23	73.76±0.26	83.33±0.34	–
		Co-Tuning + Mean Teacher	52.98±0.19	71.42±0.24	75.38±0.29	–
		Co-Tuning + FixMatch	42.34±0.19	73.24±0.25	83.13±0.34	–
		Self-Tuning (ours)	**72.50**±0.45	**83.58**±0.28	**88.11**±0.29	**90.67**±0.23

Experiments

Dataset	Type	Method	Label Proportion			
			15%	30%	50%	100%
FGVC Aircraft	TL	Fine-tuning (baseline)	39.57±0.20	57.46±0.12	67.93±0.28	81.13±0.21
		L^2-SP (Li et al., 2018)	39.27±0.24	57.12±0.27	67.46±0.26	80.98±0.29
		DELTA (Li et al., 2019)	42.16±0.21	58.60±0.29	68.51±0.25	80.44±0.20
		BSS (Chen et al., 2019)	40.41±0.12	59.23±0.31	69.19±0.13	81.48±0.18
		Co-Tuning (You et al., 2020)	44.09±0.67	61.65±0.32	72.73±0.08	83.87±0.09
	SSL	Π-model (Laine & Aila, 2017)	37.32±0.25	58.49±0.26	65.63±0.36	–
		Pseudo-Labeling (Lee, 2013)	46.83±0.30	62.77±0.31	73.21±0.39	–
		Mean Teacher (Tarvainen & Valpola, 2017)	51.59±0.23	71.62±0.29	80.31±0.32	–
		UDA (Xie et al., 2020)	43.96±0.45	64.17±0.49	67.42±0.53	–
		FixMatch (Sohn et al., 2020)	55.53±0.26	71.35±0.35	78.34±0.43	–
		SimCLRv2 (Chen et al., 2020b)	40.78±0.21	59.03±0.29	68.54±0.30	–
	Combine	Co-Tuning + Pseudo-Labeling	49.15±0.32	65.62±0.34	74.57±0.40	–
		Co-Tuning + Mean Teacher	51.46±0.25	64.30±0.28	70.85±0.35	–
		Co-Tuning + FixMatch	53.74±0.23	69.91±0.26	80.02±0.32	–
		Self-Tuning (ours)	**64.11**±0.32	**76.03**±0.25	**81.22**±0.29	**84.28**±0.14

Experiments (Unsupervised Pretrained Model)

Table 4. Classification accuracy (%) ↑ with a typical unsupervised pre-trained model MoCov2 on *CUB-200-2011*.

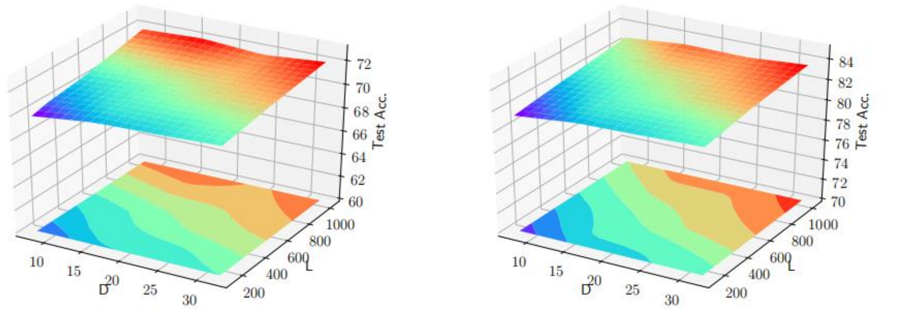
Type	Method	800 labels	5k labels
TL	Fine-Tuning (baseline)	20.04	71.50
	Co-Tuning	20.99	71.61
SSL	Mean Teacher	28.13	71.26
	FixMatch	21.18	71.28
Combine	Co-Tuning + Mean Teacher	28.43	72.21
	Co-Tuning + FixMatch	21.08	71.40
	Self-Tuning (ours)	**36.80**	**74.56**

Experiments (Ablation studies)

Table 5. Ablation studies of Self-Tuning on *Stanford Cars*.

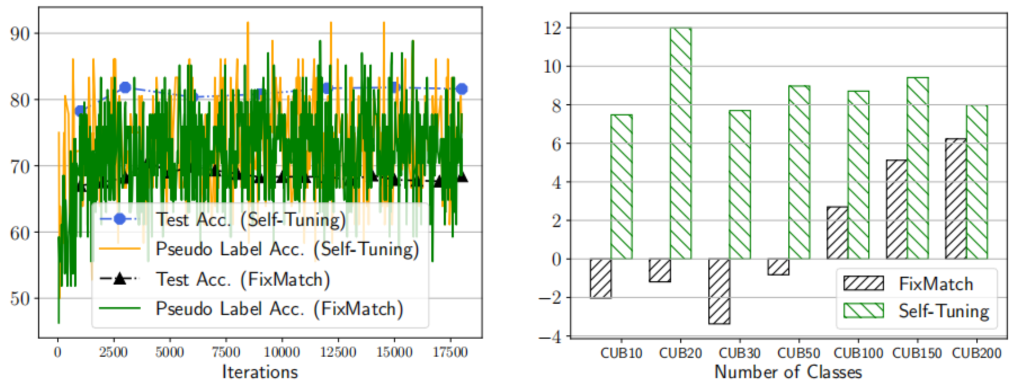
Perspective	Method	15%	30%
Loss Function	w/ CE loss	40.93	67.02
	w/ CL loss	46.29	68.82
	w/ PGC loss	**72.50**	**83.58**
Info. Exploration	w/o $\widehat{L}_{\mathrm{PGC}}$	58.82	81.71
	w/o L_{PGC}	58.85	77.52
	separate queue	70.43	80.78
	unified exploration	**72.50**	**83.58**

Experiments (Sensitivity Analysis & Others)



(a) Acc on *Car* with 15% labels (b) Acc on *Car* with 30% labels

Figure 6. Sensitivity analysis for embedded size L of the projector and queue size D of each class on *Stanford Cars*. (Warmer colors indicate higher values)



(a) Training Process on *CUB30* (b) $\mathrm{Acc}_{\mathrm{test}} - \mathrm{Acc}_{\mathrm{pseudo\_labels}}$

Figure 7. Comparisons between Self-Tuning with FixMatch on pseudo label accuracy and test accuracy.