Improving Barely Supervised Learning by Discriminating Unlabeled Data with Super-Class

Guan Gui
Nanjing University

Zhen Zhao
University of Sydney

Lei Qi
Southeast University

Luping Zhou
University of Sydney

Lei Wang
University of Wollongong

Yinghuan Shi
Nanjing University

Thirty-Sixth Conference on Neural Information Processing Systems

南京大学120周年校庆
120th ANNIVERSARY
NANJING UNIVERSITY
1902 - 2022

CONTENT

01

Introduction

02

Method

03

Experiments

南京大学120周年校庆
120th ANNIVERSARY
NANJING UNIVERSITY
1902 - 2022

01 Introduction

Why SSL models fails in barely supervised learning?

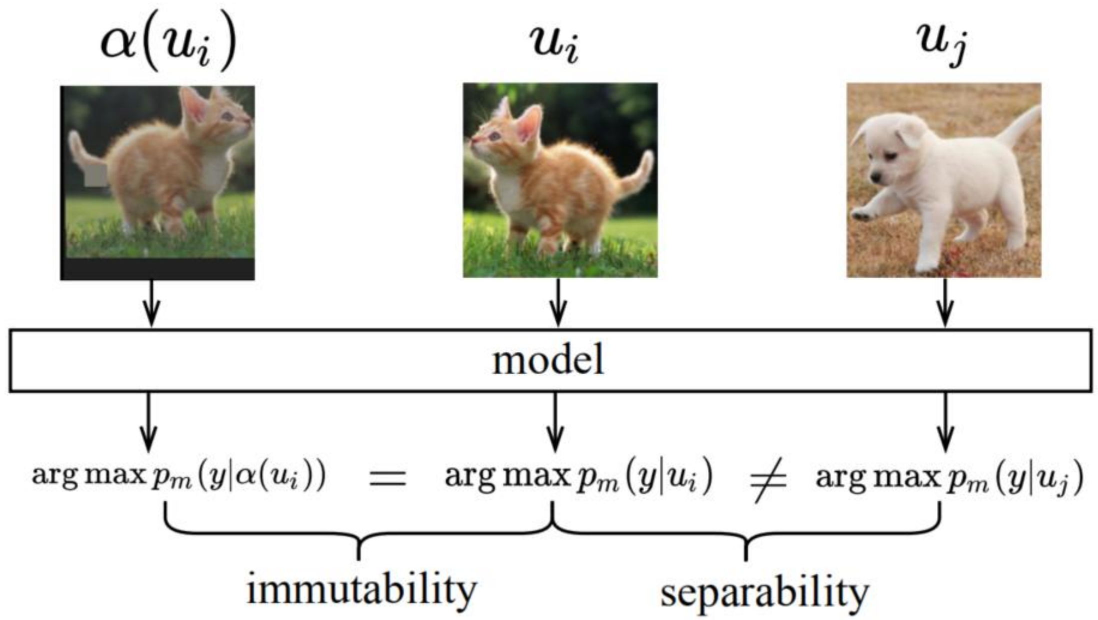
Investigate Classification Models From Immutability and Separability

- **Immutability**

The capacity of the model to be robust to perturbations.

- **Separability**

The capacity of the model to differentiate two different categories of samples.



How an semi-supervised learning model learns immutability and separability?

Scarce Discriminative Information Learning in Barely Supervised learning

- **Immutability: learning consistent information (unlabeled data)**

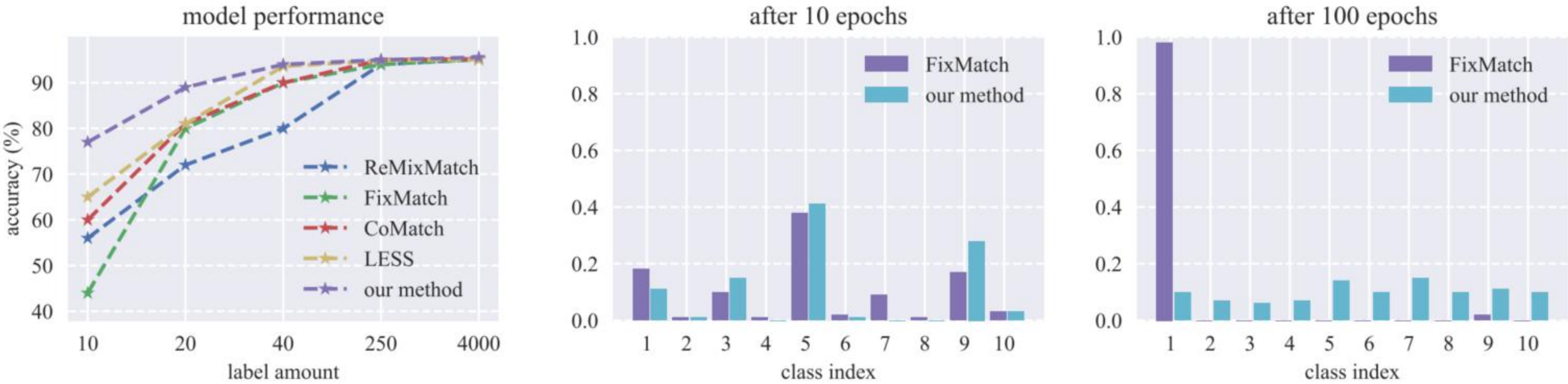
$$\ell_s = \frac{1}{B}\sum_{b=1}^{B} \mathrm{H}(p_b, p_{\mathrm{m}}(y \mid \alpha(x_b)))$$

- **Separability: learning discriminative information (labeled data)**

$$\ell_u = \frac{1}{\mu B}\sum_{b=1}^{\mu B} \mathbb{1}(\max(q_b) \geq \tau)\, \mathrm{H}(\hat{q}_b, p_{\mathrm{m}}(y \mid \mathcal{A}(u_b)))$$

- **Barely supervised learning (BSL)**

Scarce labeled data is not sufficient to provide sufficient discriminative information, resulting in the failure of the SSL model.

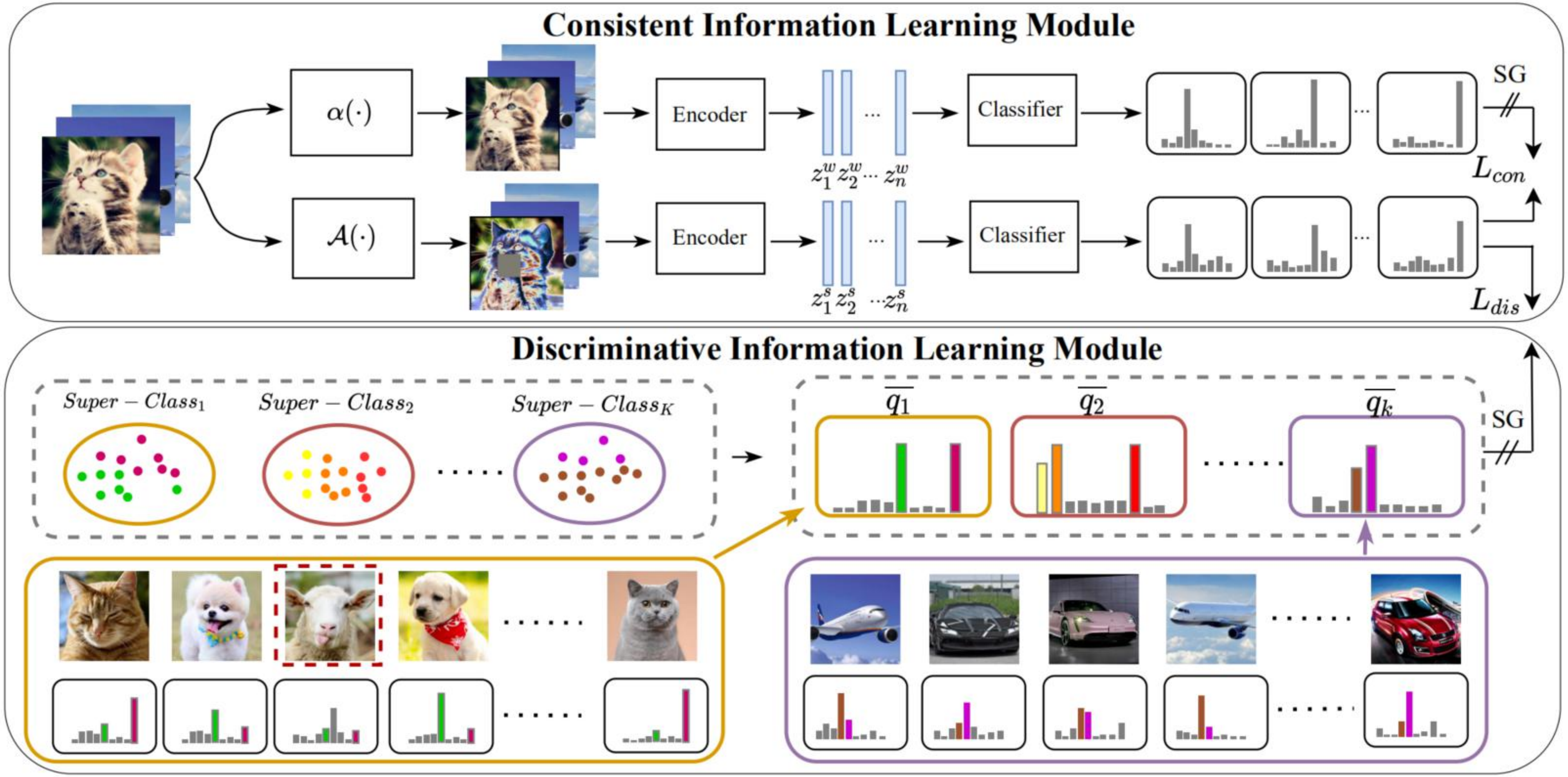


南京大学120周年校庆
120th ANNIVERSARY
NANJING UNIVERSITY
1902 - 2022

02 Method

How to mine discriminative information from unlabeled data?

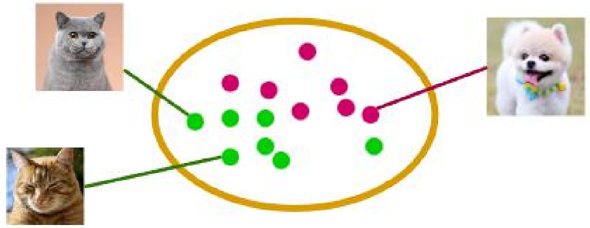
A Novel Discriminative Information Learning Module



Learning Similarity Relationships Between Samples and Super-classes

Super-class: a coarse cluster

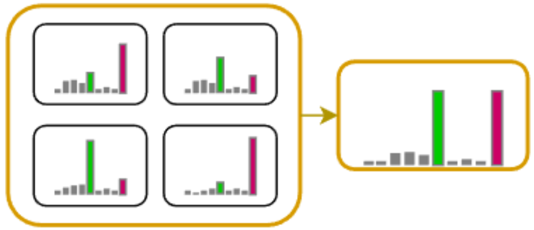
We cluster unlabeled samples into K clusters at the feature space, each cluster will have samples from multiple categories at the same time.



Super-class representation: the character of the categories

We calculate the mean of the predicted probability distribution for all samples and use this as a representation of the super-class.

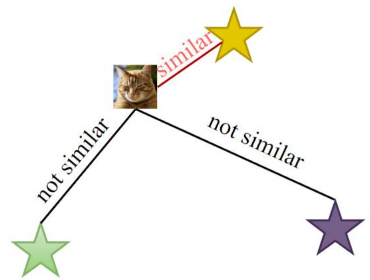
$$\overline{q_k} = \frac{1}{|C_k|}\sum_{i=1}^{|C_k|} p_m(y|\alpha(u_i)), \quad \text{with } u_i \in C_k$$



Discriminative distribution loss

we design a contrastive-like distribution loss to distinguish the sample from other super-classes.

$$L_{dis} = -\frac{1}{|B_u|}\sum_{i=1}^{|B_u|} \mathbb{1}(\max(p_m(y|\alpha(u_i))) \geq \tau_2) \log \frac{\exp(p_m(y|\mathcal{A}(u_i)) \cdot \overline{q_k}/T)}{\sum_{j=1}^{K} \exp(p_m(y|\mathcal{A}(u_i)) \cdot \overline{q_j}/T)}$$



Discriminative Information: From Coarse to Fine

Super-class: simplify the clustering task to simplify the clustering task

- Ideally, samples of the same category will form a separate cluster so that the model can discriminate the samples from all other clusters of samples. However, forming **such fine-grained clusters carries a considerable risk of errors**, especially for tasks with a large number of object categories.
- Instead of fine-grained clusters, we simplify the clustering task by allowing a cluster to contain multiple categories. In this way, the discriminative information is relatively **weakened but more robust to clustering errors**.

Progressive super-class construction: from coarse to fine information

- When the model is not well trained at the beginning, we use a small K to form the coarser super-classes to ease the clustering task and thus attain relatively reliable discriminative guidance.
- When the model is better trained, to avoid the training of the model being stagnant due to the limitation of discriminative information, we gradually increase K to provide enhanced discriminative guidance.
- We adopt a linear-step growth strategy to adjust K dynamically:

$$K = K_i, \quad \text{if} \quad K_i \leq \frac{t}{\alpha * t_s} < K_{i+1}$$

03 Experiments

Significant model performance improvements

Significant Model Performance Improvements

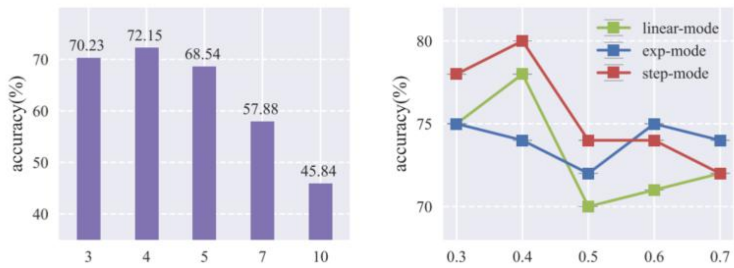
- **Significantly better than other SSL and BSL models**

Method	CIFAR-10 10 labels	CIFAR-10 20 labels	CIFAR-100 100 labels	CIFAR-100 200 labels	STL-10 10 labels	STL-10 20 labels
Mean-Teacher	15.48 ± 3.19	17.50 ± 1.16	5.17 ± 2.52	8.26 ± 3.43	11.05 ± 6.45	15.99 ± 6.45
MixMatch	17.18 ± 4.45	26.45 ± 8.17	12.85 ± 2.21	21.56 ± 4.84	10.94 ± 5.18	21.48 ± 3.17
ReMixMatch	60.29 ± 15.20	78.56 ± 9.63	26.18 ± 3.79	35.90 ± 3.66	30.86 ± 10.80	45.58 ± 8.36
FixMatch	44.47 ± 24.99	80.46 ± 5.15	25.49 ± 4.37	35.55 ± 1.59	25.75 ± 8.99	48.98 ± 6.46
FixMatch (w/DA)	67.79 ± 15.42	84.16 ± 9.27	31.10 ± 2.29	43.22 ± 1.87	42.08 ± 6.24	54.76 ± 5.44
CoMatch	60.79 ± 12.42	81.19 ± 8.55	27.54 ± 4.25	36.98 ± 2.17	29.11 ± 9.31	50.20 ± 7.57
FlexMatch	66.07 ± 10.58	85.69 ± 6.24	31.50 ± 3.61	38.05 ± 2.66	41.17 ± 6.20	54.30 ± 5.65
SLA	65.87 ± 10.83	81.89 ± 6.77	28.45 ± 2.16	38.65 ± 2.67	32.38 ± 8.32	47.50 ± 6.38
LESS	64.40 ± 10.90	81.20 ± 5.60	28.20 ± 3.00	42.50 ± 3.20	34.25 ± 7.19	48.98 ± 5.19
our method	$\mathbf{76.76 \pm 6.78}$	$\mathbf{88.49 \pm 3.26}$	$\mathbf{37.50 \pm 1.72}$	$\mathbf{45.62 \pm 1.39}$	$\mathbf{52.51 \pm 3.20}$	$\mathbf{57.98 \pm 3.18}$

- **More stable**

seed	1	2	3	4	5
FixMatch	19.15	85.11	52.52	17.09	48.50
our method	**81.28**	**86.12**	**70.34**	**74.90**	**71.17**

- **Insensitive to K**



(a) K (b) α

Thanks for Listening