南京航空航天大學

Nanjing University of Aeronautics and Astronautics

Exploring Balanced Feature Spaces for Representation Learning

Bingyi Kang[1], Yu Li [2], Zehuan Yuan[3], Jiashi Feng[1]

[1]National University of Singapore, [2]Institute of Computing Technology, CAS, [3]ByteDance AI Lab

kang@u.nus.edu,liyu@ict.ac.cn,yuanzehuan@bytedance.com,elefjia@nus.edu.sg

ICLR 2021

Introduction

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

$$\mathcal{L}_{\mathrm{CE}} = \frac{1}{N}\sum_{i=1}^{N} -\log p_{y_i},$$

y_i is supervision signal.

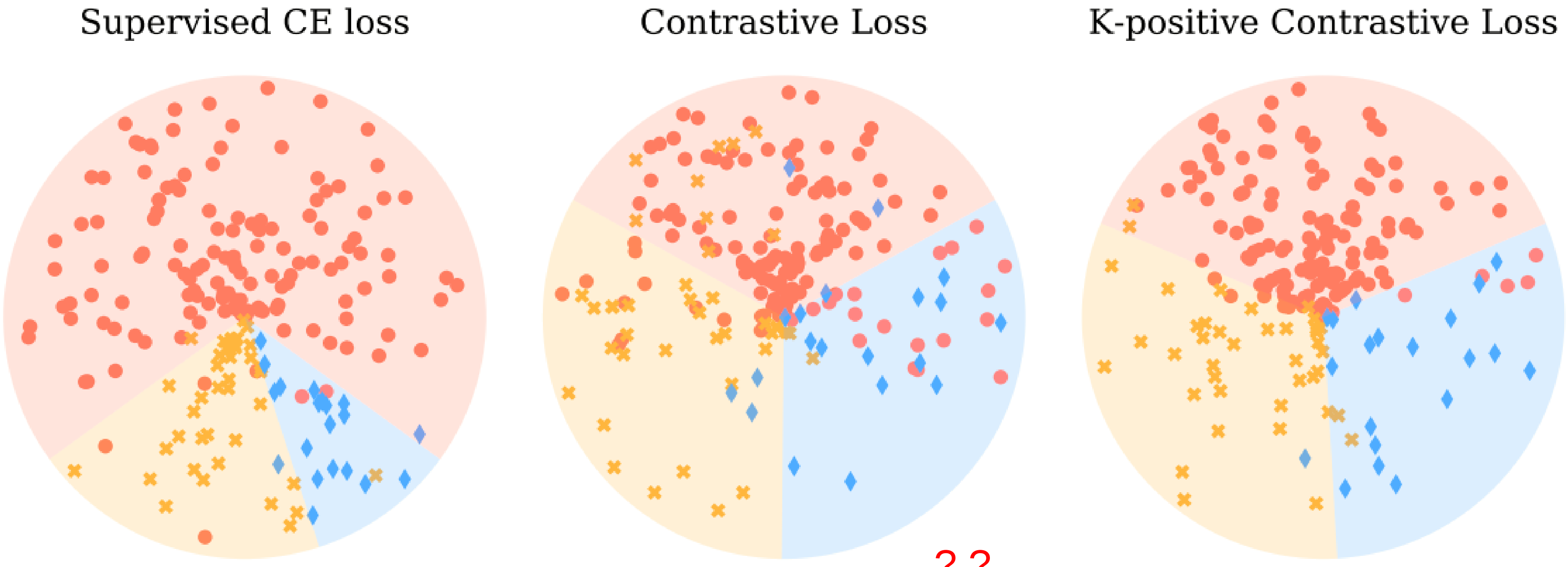
v_i^+ typically produced by data augmentation.

$$\mathcal{L}_{\mathrm{CL}} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\exp(v_i \cdot v_i^+ / \tau)}{\exp(v_i \cdot v_i^+ / \tau) + \sum_{v_i^- \in V^-} \exp(v_i \cdot v_i^- / \tau)},$$

No supervision signal.

Introduction

- Feature spaces learned with different losses on **long-tailed** datasets.



bised but **discriminative** | reverse ⇄ | **balanced** but not discriminative ? ? | **balanced** and **discriminative** ?

- Whether the feature spaces learned by unsupervised **C**ontrastive **L**oss (**CL**) is **balanced**?
- Benefits of balanced feature spaces?
- How to learn **balanced** and **discriminative** feature space?

Contents

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

- Whether the feature spaces learned by unsupervised Contrastive Loss is **balanced**?
- **Benefits** of balanced feature spaces?
- How to learn **balanced** and **discriminative** feature spaces?

Whether CL is balanced?

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

- Balancedness metric

A feature space **V** is balanced if the representations {vi} from different classes within it have similar degrees of linear separability.

the accuracy of a linear classifier

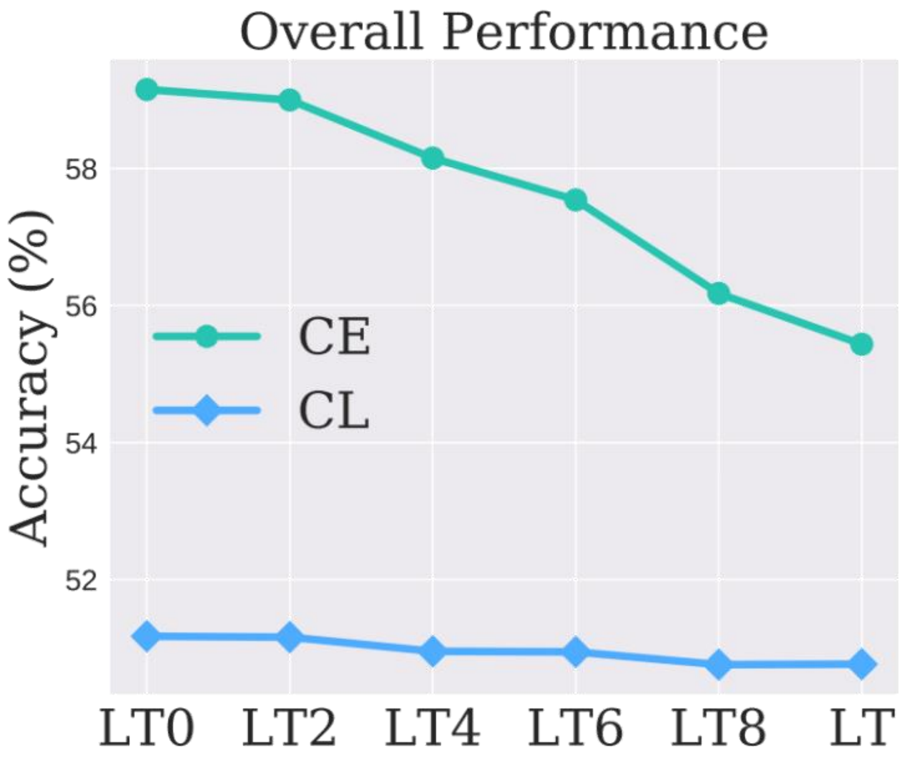
$$\beta(V) \triangleq \frac{1}{C^2}\sum_{i,j}^{C}\exp\left(-\frac{|a_i-a_j|^2}{\sigma}\right), \text{ where } a_j=\frac{\#\{v_i|\hat{y}_i=j, y_i=j, v_i\in V\}}{\#\{v_i|y_i=j, v_i\in V\}}. \quad (3)$$

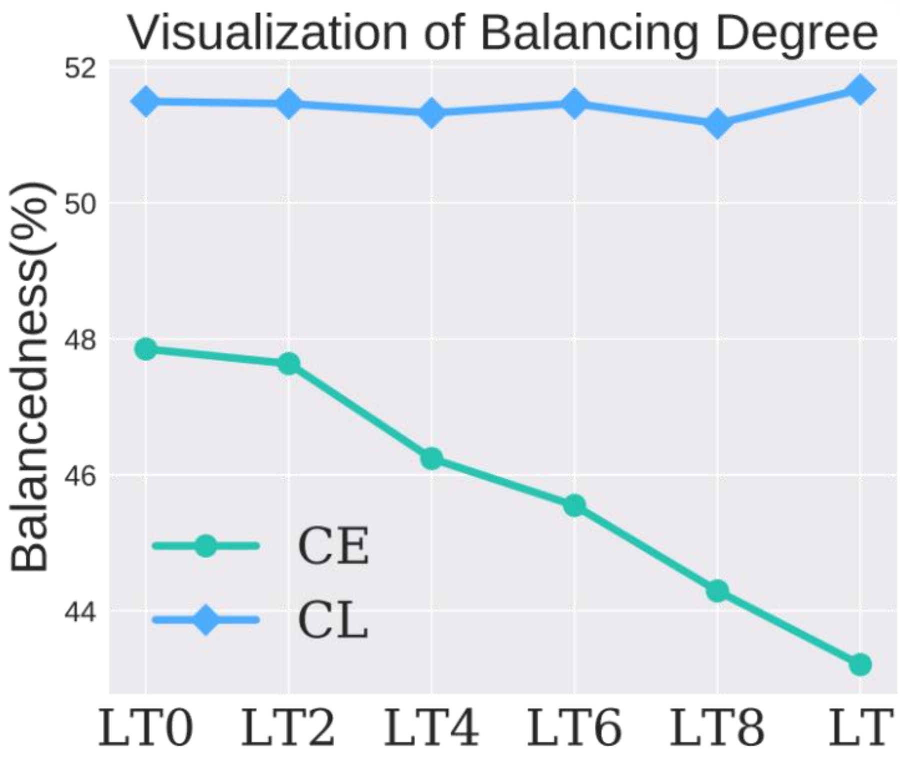
C is number of classes, σ is a fixed parameter.

This metric achieves its maximum when all the class-wise accuracies are equal, i.e., there being no separability bias of the learned representations to any class.

$\boxed{\text{metric } \beta} \propto \boxed{\text{Balanced}}$

Whether CL is balanced?





more imbalanced

more imbalanced

- The model trained with the **unsupervised contrastive loss** generates a more **balanced** feature space.

What are the **benefits** from a **balanced** model for recognition?

Contents

- Whether the feature spaces learned by unsupervised Contrastive Loss is **balanced**?
- **Benefits** of balanced feature spaces?
- How to learn **balanced** and **discriminative** feature spaces?

Benefits of balanced feature spaces?

- Out-of-distribution generalization (open-set)
- Cross-domain and cross-task generalization

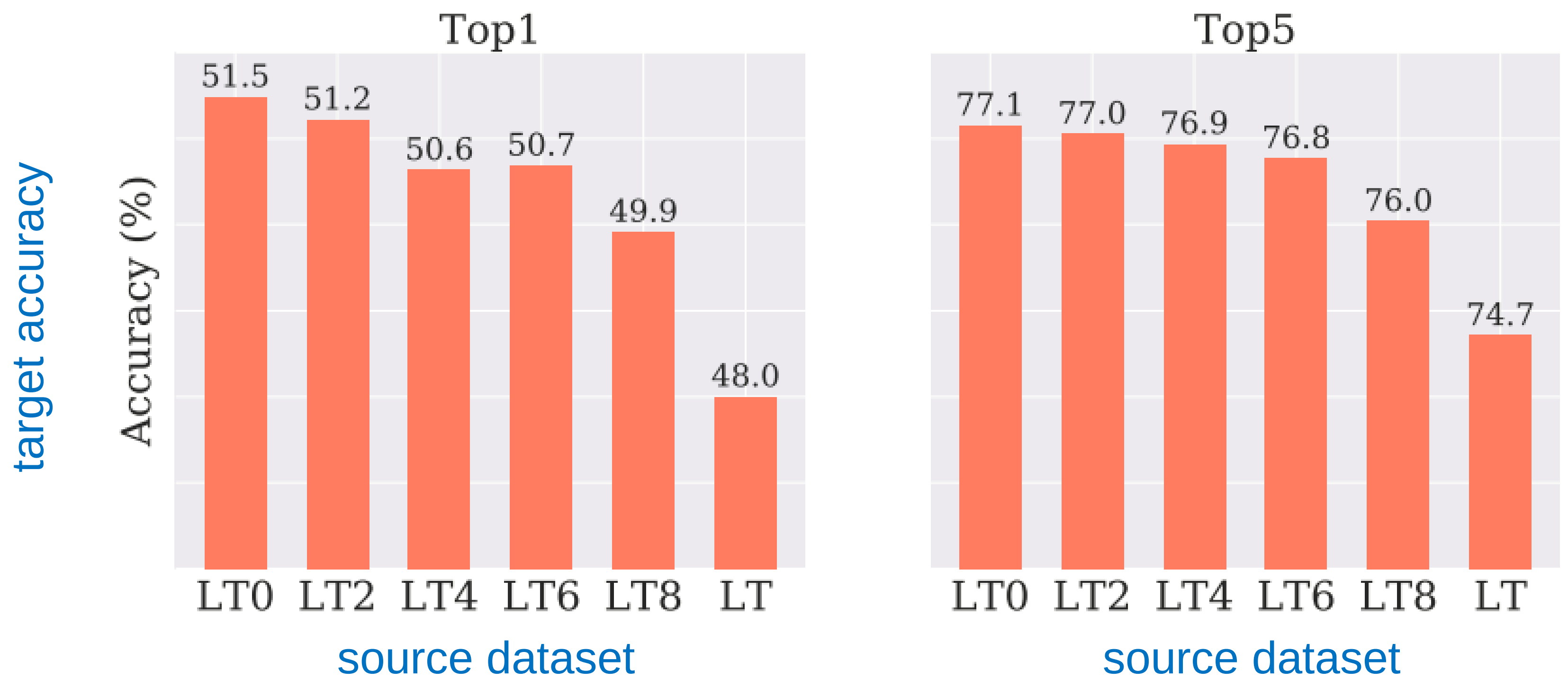
Benefits of balanced feature spaces?

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

- Out-of-distribution generalization (open-set)
- Cross-domain and cross-task generalization

Balance VS generalization?

CL ×
CE √



- More **balanced** representation models tend to **generalize better** for recognizing unseen classes.

Benefits of balanced feature spaces?

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

- Out-of-distribution generalization (open-set)
- Cross-domain and cross-task generalization

Table 1: Results on Places365, VOC and COCO. AP_{50} is the default metric for VOC, while AP^{bb} and AP^{mk} denote the bounding-box and mask AP for COCO respectively. Black / gray numbers correspond to results of the representation models trained on ImageNet-LT / ImageNet respectively. See appendix for complete results.

	cross-domain	*cross-task*		
	Places365 (Top1)	VOC (AP_{50})	COCO (AP^{bb})	COCO (AP^{mk})
CE	38.50 / 46.06	76.45 / 81.26	38.13 / 40.08	33.29 / 34.85
CL	41.24 / 46.16	78.19 / 82.28	39.67 / 40.41	34.73 / 35.14
$\Delta_{CL,CE}$	**+2.74** / +0.10	**+1.64** / +0.02	**+1.54** / +0.33	**+1.44** / +0.29

- The generalization performance not simply stem from using **self-supervised pre-training**, but indeed come from learning more **balanced** feature spaces.

Benefit : generalization

Contents

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

- Whether the feature spaces learned by unsupervised Contrastive Loss is **balanced**?
- **Benefits** of balanced feature spaces?
- How to learn **balanced** and **discriminative** feature spaces?

Introduction

$$\mathcal{L}_{\mathrm{CE}} = \frac{1}{N}\sum_{i=1}^{N} -\log p_{y_i},$$

y_i is supervision signal.

v_i^+ typically produced by data augmentation.

$$\mathcal{L}_{\mathrm{CL}} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\exp(v_i \cdot v_i^+/\tau)}{\exp(v_i \cdot v_i^+/\tau) + \sum_{v_i^- \in V^-}\exp(v_i \cdot v_i^-/\tau)},$$

No supervision signal.

Introduction

$$\mathcal{L}_{\mathrm{CE}} = \frac{1}{N}\sum_{i=1}^{N} -\log p_{y_i},$$

y_i is supervision signal.

v_i^+ typically produced by data augmentation.

$$\mathcal{L}_{\mathrm{CL}} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\exp(v_i \cdot v_i^+/\tau)}{\exp(v_i \cdot v_i^+/\tau) + \sum_{v_i^- \in V^-} \exp(v_i \cdot v_i^-/\tau)},$$

No supervision signal.

Supervised **c**ontrastive **l**earning (**SCL**) :

- **SCL** uses **all** the instances from the same class to construct the positive pairs, which cannot avoid the dominance of head classes.

K-POSITIVE CONSTRASTIVE LOSS

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Contrastive **L**earning have **limited** capability of **semantic discrimination,** enen two instances from the **same** class are forced to be apart from each other in the learned feature space.

$$\mathcal{L}_{\mathrm{KCL}} = \frac{1}{N(k+1)} \sum_{i=1}^{N} \sum_{v_j^+ \in \{\tilde{v}_i\} \cup V_{i,k}^+} - \log \frac{\exp(v_i \cdot v_j^+ / \tau)}{\exp(v_i \cdot \tilde{v}_i / \tau) + \sum_{v_j \in V_i} \exp(v_i \cdot v_j / \tau)}, \qquad (4)$$

Draws ***k*** instances from the same class to form the **positive sample set** $V_{i,k}^+$, instead of only using its augmentation.

k brings two benefits:

- It uses the label information to learn **discriminative** representations.
- It uses the same number of instances (i.e., k) for all the classes, which **balances** the representations.

The difference with **s**upervised **c**ontrastive **l**earning (**SCL**) :

- **SCL** uses **all** the instances from the same class to construct the positive pairs, which cannot avoid the dominance of head classes.

Contributions

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

- We are the **first** to study self-supervised contrastive learning on imbalanced datasets.
- We are the **first** to reveal that the model trained by contrastive learning can learn **balanced** feature spaces.
- Our empirical analysis proves that learning **balanced** feature spaces **benefits** the **generalization** of representation models.
- We develop the k-positive contrastive learning (**KCL**) method to learn **balanced** and **discriminative** feature representations.

Experiments

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

- Experiments on **ImageNet-LT**(left) and **iNaturalist**(right).

Table 2: ImageNet-LT results

Method	Many	Medium	Few	All
OLTR (Liu et al., 2019)[a]	35.8	32.3	21.5	32.2
Joint (SL-i) (Kang et al., 2020)	64.9	35.2	6.8	42.5
τ-norm (Kang et al., 2020)	56.6	44.2	27.4	46.7
cRT (Kang et al., 2020)	58.8	44.0	26.1	47.3
FCL	61.4	47.0	28.2	49.8
KCL	61.8	49.4	30.9	**51.5**

[a] Reproduced by re-running their code with ResNet50.

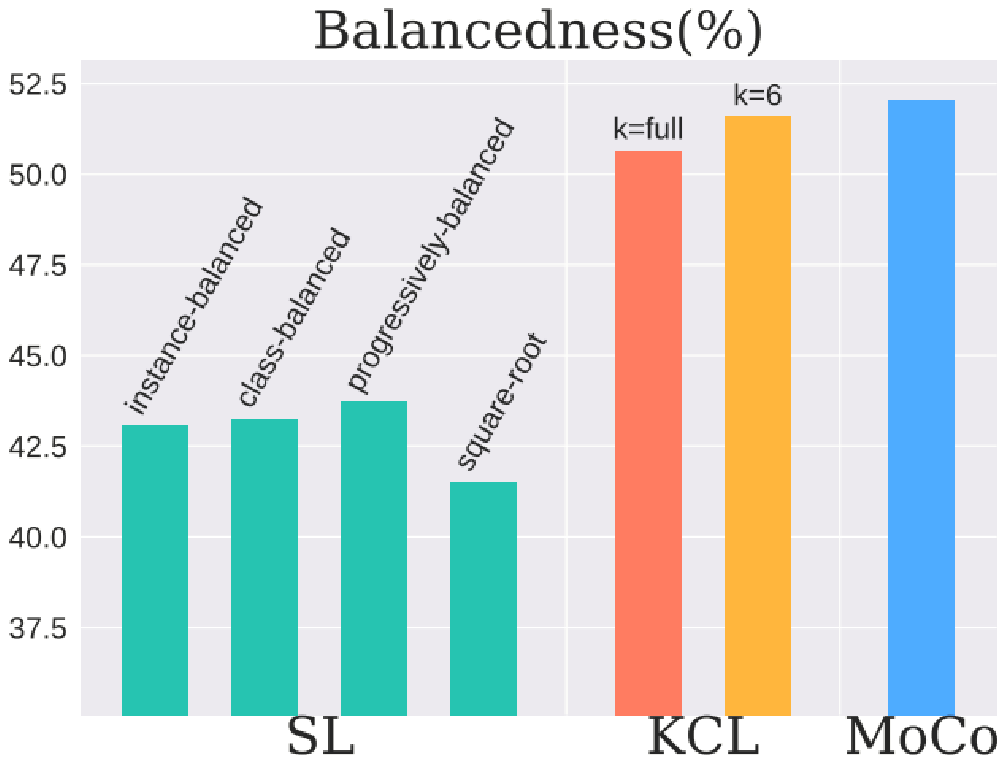
Table 3: iNaturalist 2018 results

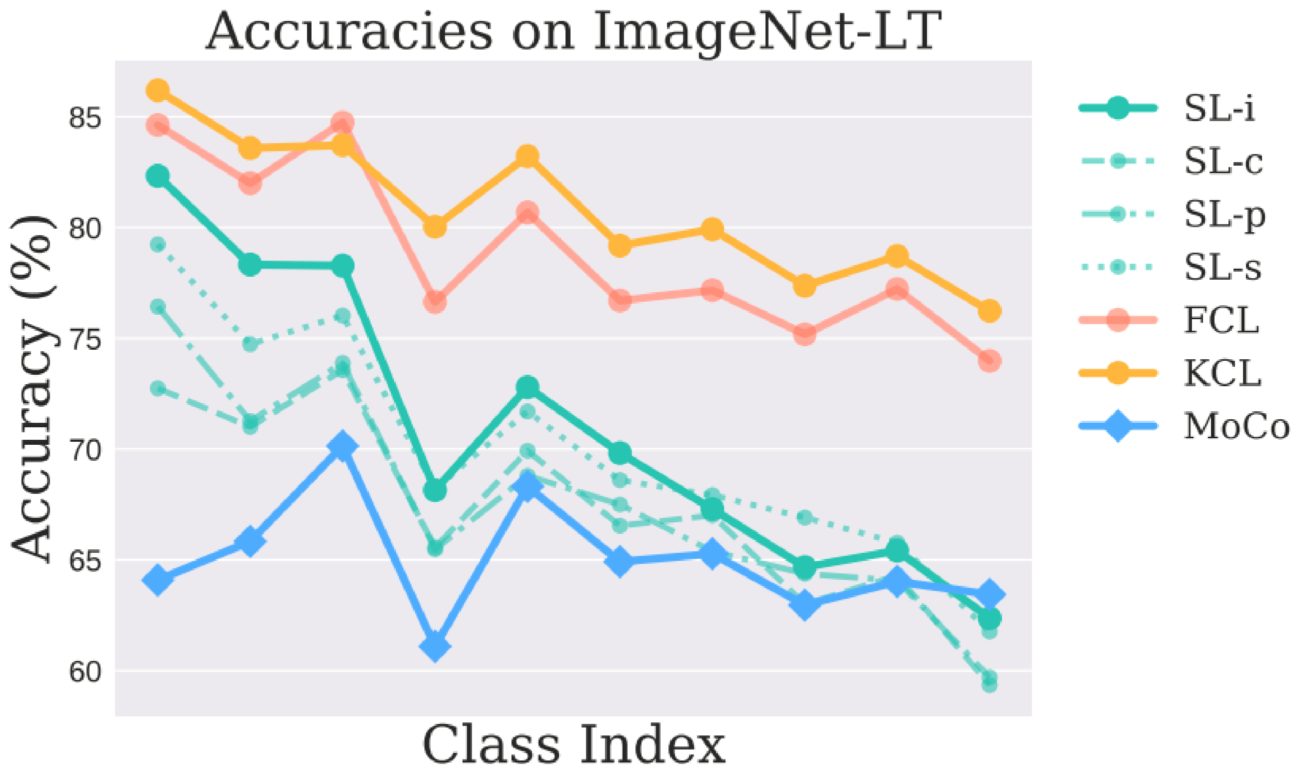
Method	Top1
CB-Focal (Cui et al., 2019)	61.1
LDAM (Cao et al., 2019)	64.6
LDAM+DRW (Cao et al., 2019)	68.0
cRT (Kang et al., 2020)	65.2
τ-norm (Kang et al., 2020)	65.6
BBN (Zhou et al., 2020)	66.3
FCL	66.4
KCL	**68.6**

Experiments

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

- Experiments on feature space **balancedness**(left) and **class-wise accuracy**.





Experiments

Pre-training representation models for **downstream** tasks.

- Out-of-distribution (OOD) Generalization(Open-set).
- Cross-domain and cross-task generalization.

Table 4: *OOD generalization results (top-1 accuracy) on balanced datasets.* We use the *source* classes of origninal ImageNet to learn representation network (ResNet50), and use the *target* classes and *all* classes respectively to learn linear classifiers for evaluation.

	Split-overlap			*Split-independent*		
	source	target	all	source	target	all
CE	81.2	70.7	67.2	82.8	50.3	62.4
CL	67.0	60.1	58.3	68.2	54.8	58.2
KCL	81.4	74.8 (+4.1)	70.8 (+3.6)	83.2	58.1 (+3.3)	67.2 (+4.8)

Table 5: Comparison of different representation learning methods for the downstream tasks.

	repr	VOC (AP_{50})	COCO (AP^{bb})	COCO (AP^{mk})
SL	76.6	81.26	40.08	34.85
MoCo	60.6	81.28	40.41	35.15
KCL	76.8	**82.32**	**40.79**	**35.45**

THANKS