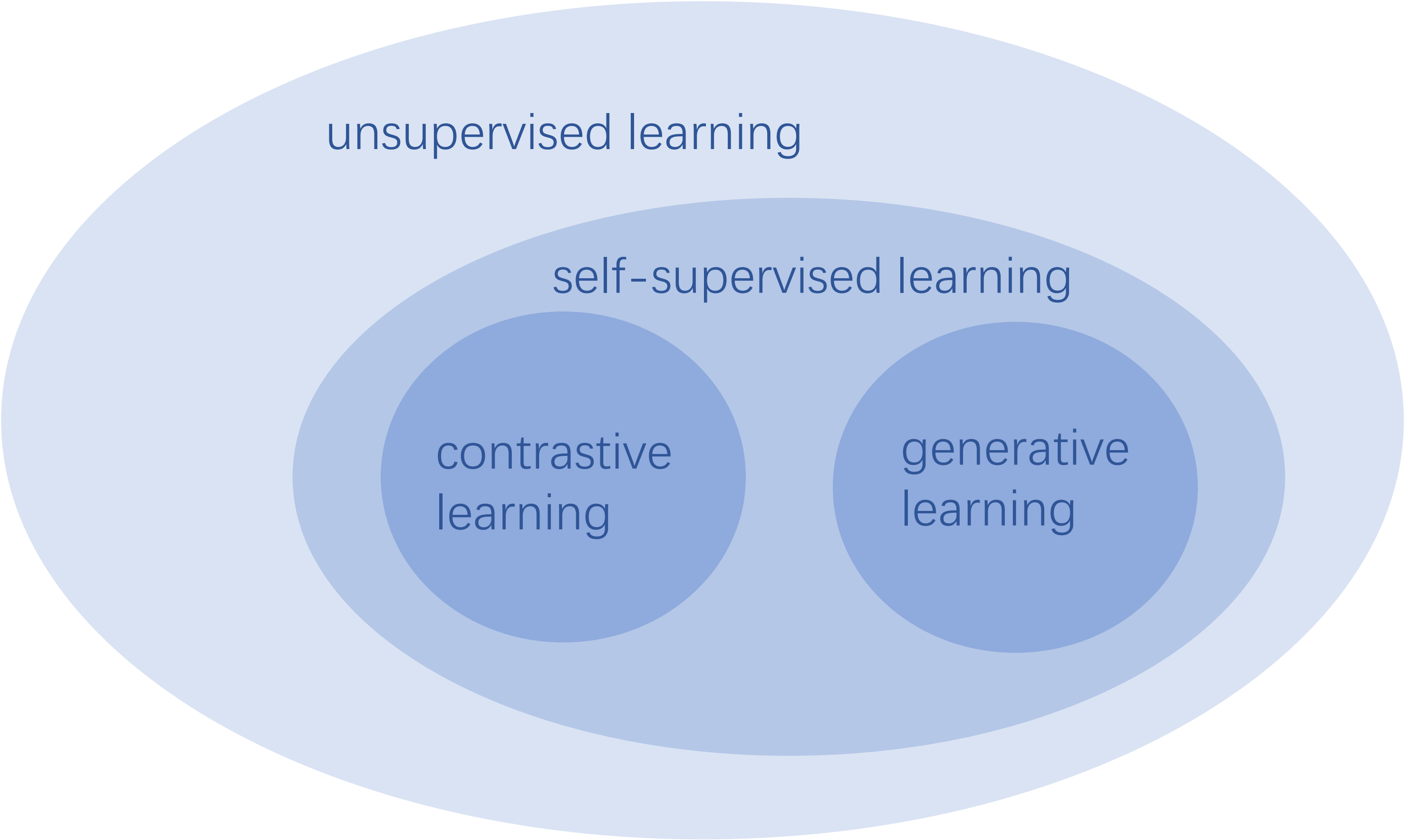
CO2: CONSISTENT CONTRAST FOR UNSUPERVISED VISUAL REPRESENTATION LEARNING

Chen Wei, Huiyu Wang, Wei Shen, Alan Yuille
Johns Hopkins University
{weichen3012,williamwanghuiyu,shenwei1231,alan.l.yuille}@gmail.com

ICLR2021

contrastive learning

Introduction



Self-Supervised learning

A form of unsupervised learning where the **data** provides the supervision

(credit to Andrew Zisserman)

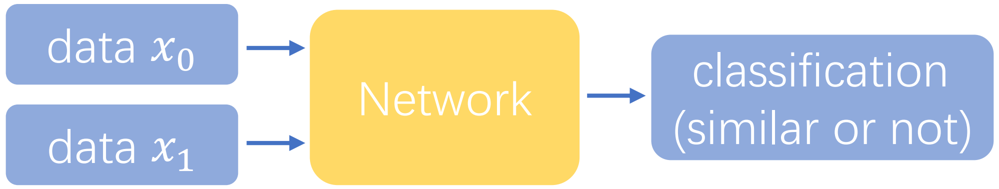
https://project.inria.fr/paiss/files/2018/07/zisserman-self-supervised.pdf

Introduction

- contrastive learning
  - positive samples + negative samples
  - loss in feature space
  - difficulty: choice of positive/negative samples
- generative learning
  - encoder-decoder
  - pixel-wise loss
  - difficulty: pixel–wise reconstruction

Introduction

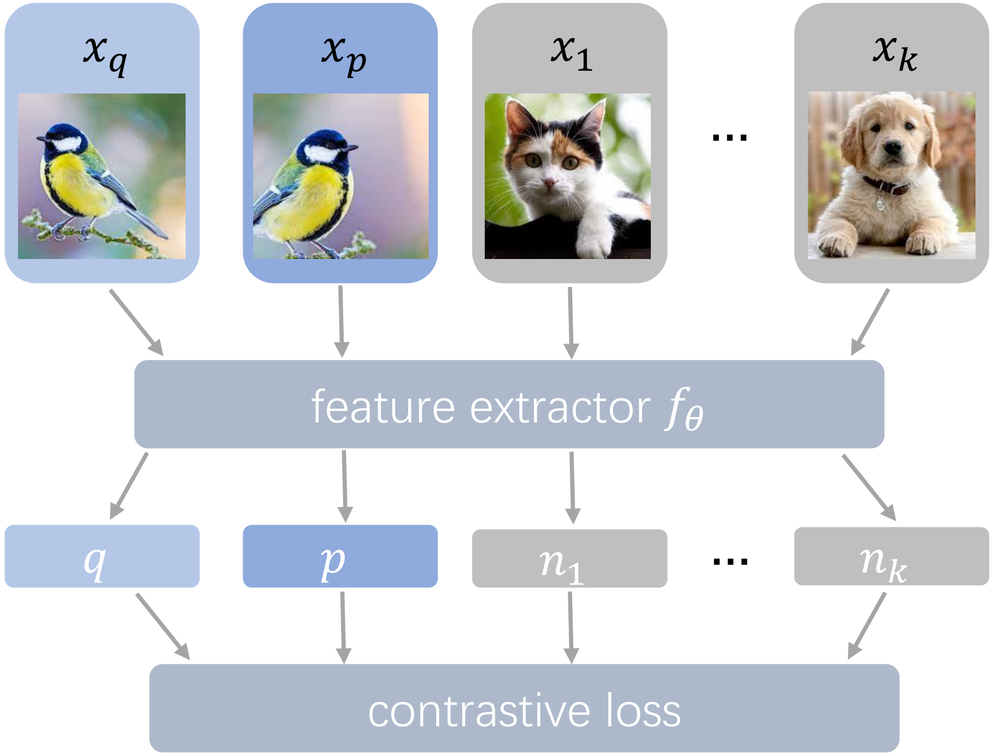
- contrastive learning
  - positive samples + negative samples
  - loss in feature space
  - difficulty: choice of positive/negative samples
- generative learning
  - encoder-decoder
  - pixel-wise loss
  - difficulty: pixel–wise reconstruction



Contrastive Learning

- x_q query
- x_p positive sample
- $x_1, \ldots, x_k$ negative samples
- f_θ feature extractor (trained on positive/negative samples)
- $p = f_\theta(x_p)$
- $q = f_\theta(x_q)$
- $n_i = f_\theta(x_i), i = 1, \ldots, k$
- InfoNCE loss:

$$\mathcal{L}_{ins} = -\log \frac{\exp(\mathbf{q}\cdot\mathbf{p}/\tau_{ins})}{\exp(\mathbf{q}\cdot\mathbf{p}/\tau_{ins}) + \sum_{k=1}^{K}\exp(\mathbf{q}\cdot\mathbf{n}_k/\tau_{ins})}$$

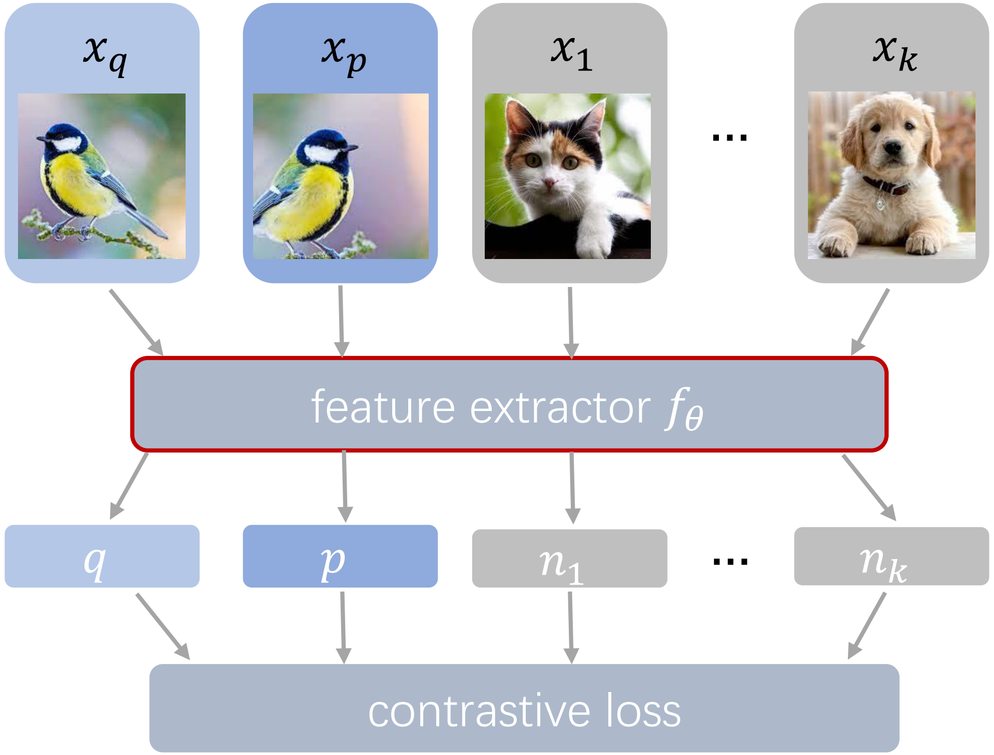


Contrastive Learning

Train the feature extractor,
then test downstream tasks using feature extractor

- x_q query
- x_p positive sample
- $x_1, \dots, x_k$ negative samples
- f_θ feature extractor (trained on positive/negative samples)
- $p = f_\theta(x_p)$
- $q = f_\theta(x_q)$
- $n_i = f_\theta(x_i), i = 1, \dots, k$
- InfoNCE loss:

$$\mathcal{L}_{ins} = -\log \frac{\exp(\mathbf{q} \cdot \mathbf{p} / \tau_{ins})}{\exp(\mathbf{q} \cdot \mathbf{p} / \tau_{ins}) + \sum_{k=1}^{K} \exp(\mathbf{q} \cdot \mathbf{n}_k / \tau_{ins})}$$



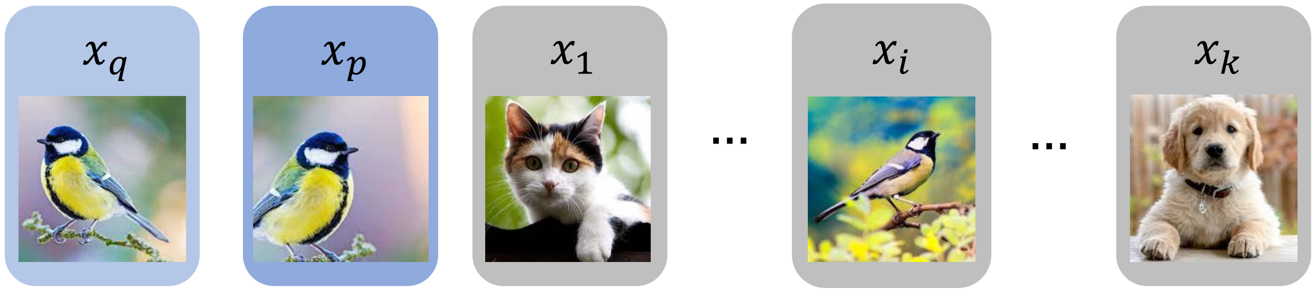
Motivation

- in InfoNCE loss:

$$\mathcal{L}_{ins} = -\log \frac{\exp(\mathbf{q}\cdot\mathbf{p}/\tau_{ins})}{\exp(\mathbf{q}\cdot\mathbf{p}/\tau_{ins}) + \sum_{k=1}^{K}\exp(\mathbf{q}\cdot\mathbf{n}_k/\tau_{ins})}$$

  use “zero-one” label

- but these “hard negative” crops in fact tend to be semantically close



Motivation

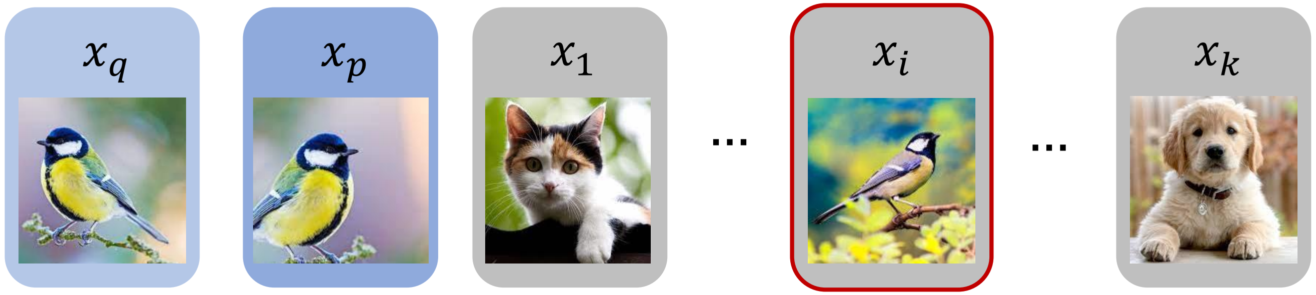
- in InfoNCE loss:

$$\mathcal{L}_{ins} = -\log \frac{\exp(\mathbf{q}\cdot\mathbf{p}/\tau_{ins})}{\exp(\mathbf{q}\cdot\mathbf{p}/\tau_{ins}) + \sum_{k=1}^{K}\exp(\mathbf{q}\cdot\mathbf{n}_k/\tau_{ins})}$$

-> to built a consistent contrast

use “zero-one” label

- but these “hard negative” crops in fact tend to be semantically close



heterogeneous similarity

Method

- Make features p and q have consistent distance from $n_i, i = 1, \ldots, k$

$$\mathcal{L} = \mathcal{L}_{ins} + \alpha \mathcal{L}_{con}$$

$$\mathcal{L}_{con} = \frac{1}{2} D_{\mathrm{KL}}(P \| Q) + \frac{1}{2} D_{\mathrm{KL}}(Q \| P)$$

$$P(i) = \frac{\exp(\mathbf{p} \cdot \mathbf{n}_i / \tau_{con})}{\sum_{k=1}^{K} \exp(\mathbf{p} \cdot \mathbf{n}_k / \tau_{con})}$$

$$Q(i) = \frac{\exp(\mathbf{q} \cdot \mathbf{n}_i / \tau_{con})}{\sum_{k=1}^{K} \exp(\mathbf{q} \cdot \mathbf{n}_k / \tau_{con})}$$

Method

- Make features p and q have consistent distance from $n_i, i = 1, \dots, k$

$$\mathcal{L} = \mathcal{L}_{ins} + \alpha \mathcal{L}_{con}$$

$$\mathcal{L}_{con} = \frac{1}{2} D_{\mathrm{KL}}(P \| Q) + \frac{1}{2} D_{\mathrm{KL}}(Q \| P)$$

$$P(i) = \frac{\exp(\mathbf{p} \cdot \mathbf{n}_i / \tau_{con})}{\sum_{k=1}^{K} \exp(\mathbf{p} \cdot \mathbf{n}_k / \tau_{con})}$$

P,Q can be seen as two probability distributions

$$Q(i) = \frac{\exp(\mathbf{q} \cdot \mathbf{n}_i / \tau_{con})}{\sum_{k=1}^{K} \exp(\mathbf{q} \cdot \mathbf{n}_k / \tau_{con})}$$

Method

- Make features p and q have consistent distance from $n_i, i = 1, \ldots, k$

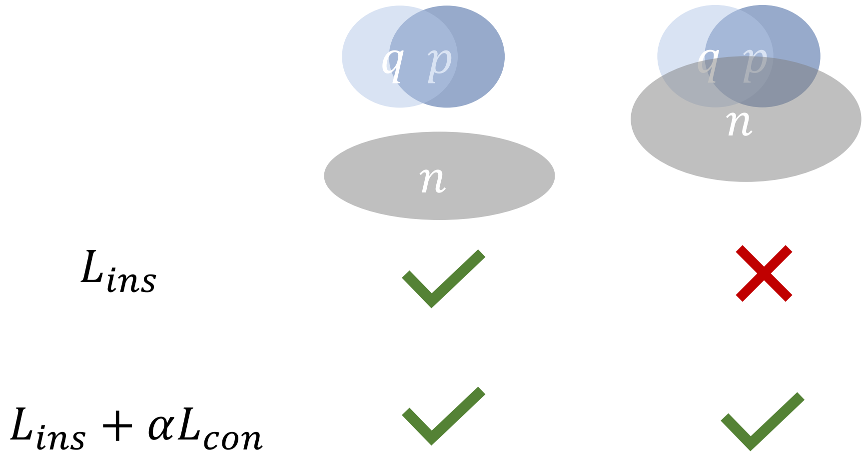
$$\mathcal{L} = \mathcal{L}_{ins} + \alpha \mathcal{L}_{con}$$

$$\mathcal{L}_{con} = \frac{1}{2} D_{\mathrm{KL}}(P \| Q) + \frac{1}{2} D_{\mathrm{KL}}(Q \| P)$$

P,Q can be seen as two probability distributions

$$P(i) = \frac{\exp(\mathbf{p} \cdot \mathbf{n}_i / \tau_{con})}{\sum_{k=1}^{K} \exp(\mathbf{p} \cdot \mathbf{n}_k / \tau_{con})}$$

$$Q(i) = \frac{\exp(\mathbf{q} \cdot \mathbf{n}_i / \tau_{con})}{\sum_{k=1}^{K} \exp(\mathbf{q} \cdot \mathbf{n}_k / \tau_{con})}$$



Experiment

- evaluate CO2 (proposed method) based on MoCo and MoCo v2

ResNet-50

- train on ImageNet-1K
- freeze the backbone network (including BN) after the unsupervised pre-training stage
- then train a supervised linear classifier (a fully-connected layer and a softmax layer) on the 2048-D features

Experiment

Table 1: Linear classification protocol on ImageNet-1K

Pretext Task	Arch.	Head	#epochs	Top-1 Acc. (%)
ImageNet Classification	R50	-	90	76.5
Exemplar (Dosovitskiy et al., 2014)	R50w3×	-	35	46.0
Relative Position (Doersch et al., 2015)	R50w2×	-	35	51.4
Rotation (Gidaris et al., 2018)	Rv50w4×	-	35	55.4
Jigsaw (Noroozi & Favaro, 2016)	R50	-	90	45.7
Methods based on contrastive learning:				
InsDisc (Wu et al., 2018)	R50	Linear	200	54.0
Local Agg. (Zhuang et al., 2019)	R50	Linear	200	58.2
CPC v2 (Hénaff et al., 2019)	$R170_w$	-	~200	65.9
CMC (Tian et al., 2019)	R50	Liner	240	60.0
AMDIM (Bachman et al., 2019)	$AMDIM_{large}$	-	150	68.1
PIRL (Misra & van der Maaten, 2020)	R50	Linear	800	63.6
SimCLR (Chen et al., 2020a)	R50	MLP	1000	69.3
MoCo (He et al., 2020)	R50	Linear	200	60.6
MoCo (He et al., 2020) + CO2	R50	Linear	200	63.5
MoCo v2 (Chen et al., 2020b)	R50	MLP	200	67.5
MoCo v2 (Chen et al., 2020b) + CO2	R50	MLP	200	68.0

Experiment

- semi-supervised learning: finetune the whole pre-trained networks with only 1% and 10% labels which are sampled in a class-balanced way

Table 2: Top-5 accuracy for semi-supervised learning on ImageNet

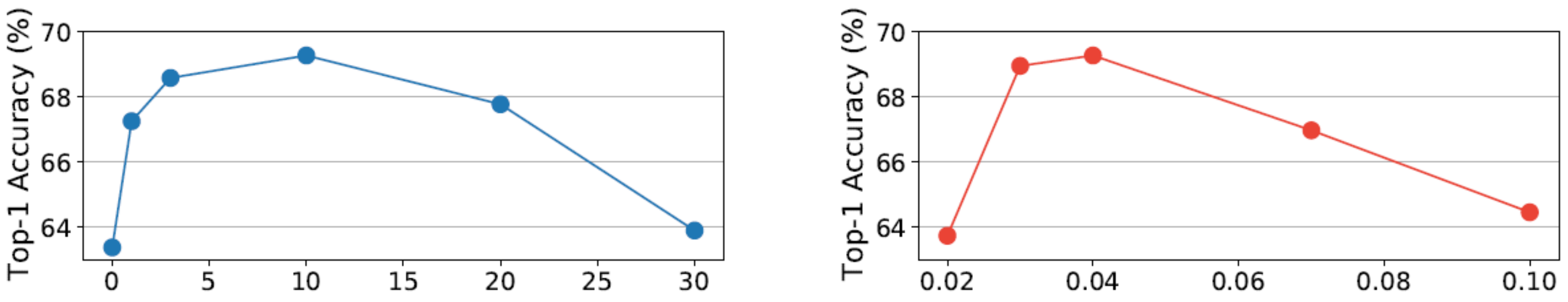
Pretext Task	1% labels	10% labels
Supervised Baseline	48.4	80.4
InsDisc (Wu et al., 2018)	39.2	77.4
PIRL (Misra & van der Maaten, 2020)	57.2	83.8
MoCo (He et al., 2020)	62.4	84.1
MoCo (He et al., 2020) + CO2	66.2	85.2
MoCo v2 (Chen et al., 2020b)	69.8	85.0
MoCo v2 (Chen et al., 2020b) + CO2	71.0	85.7

Experiment-transfer learning

Table 3: Transfer learning performance on PASCAL VOC datasets

	Image Classification	Object Detection			Semantic Segmentation
Pretext Task	mAP	AP_{50}	AP_{all}	AP_{75}	mIoU
ImageNet Classification	88.0	81.3	53.5	58.8	74.4
Rotation (Gidaris et al., 2018)	63.9	72.5	46.3	49.3	-
Jigsaw (Noroozi & Favaro, 2016)	64.5	75.1	48.9	52.9	-
InsDisc (Wu et al., 2018)	76.6	79.1	52.3	56.9	-
PIRL (Misra & van der Maaten, 2020)	81.1	80.7	54.0	59.7	-
MoCo (He et al., 2020)	-	81.5	55.9	62.6	72.5
MoCo (He et al., 2020) (*our impl.*)	79.7	81.6	56.2	62.4	72.6
MoCo (He et al., 2020) + CO2	82.6	81.9	56.0	62.6	73.3
MoCo v2 (Chen et al., 2020b)	85.0	82.4	57.0	63.6	74.2
MoCo v2 (Chen et al., 2020b) + CO2	85.2	82.7	57.2	64.1	74.7

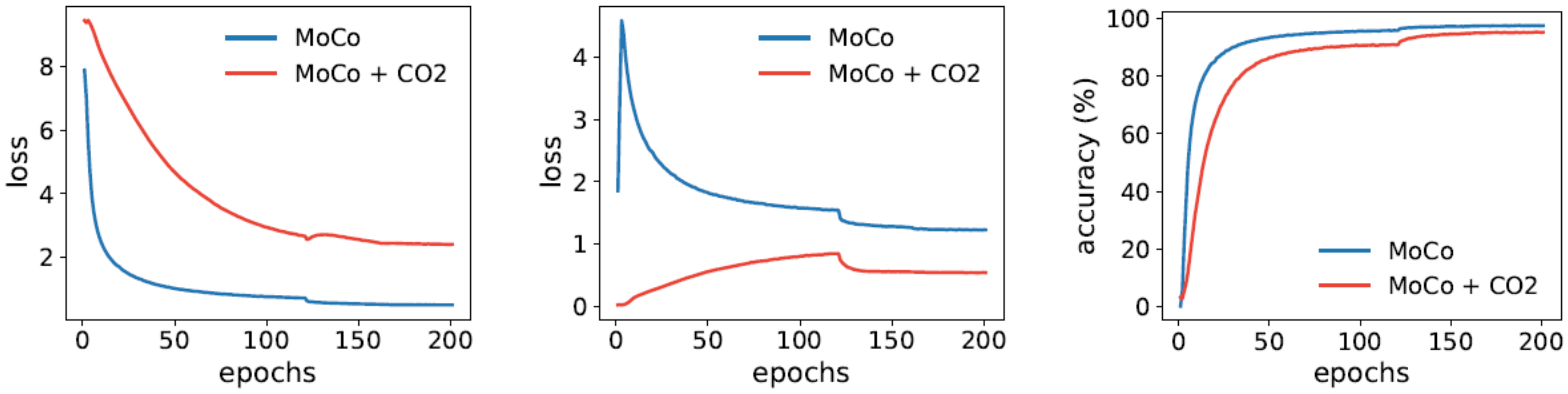
Experiment



(a) Effect of varying the coefficient α.

(b) Effect of varying the temperature τ_{con}.

Figure 2: Ablation on the effect of hyper-parameters.



(a) $\mathcal{L}_{ins}$

(b) $\mathcal{L}_{con}$

(c) Instance discrimination acc.

Figure 3: Training curves of ResNet-18 on ImageNet-100.

Discussion

- relaxes the stereotype restriction that negative labels should always be known and clean
- easily applied to other contrastive learning mechanisms

- it is an example of similarity of feature
- but for contrastive learning, choice of positive/negative samples are more important