上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

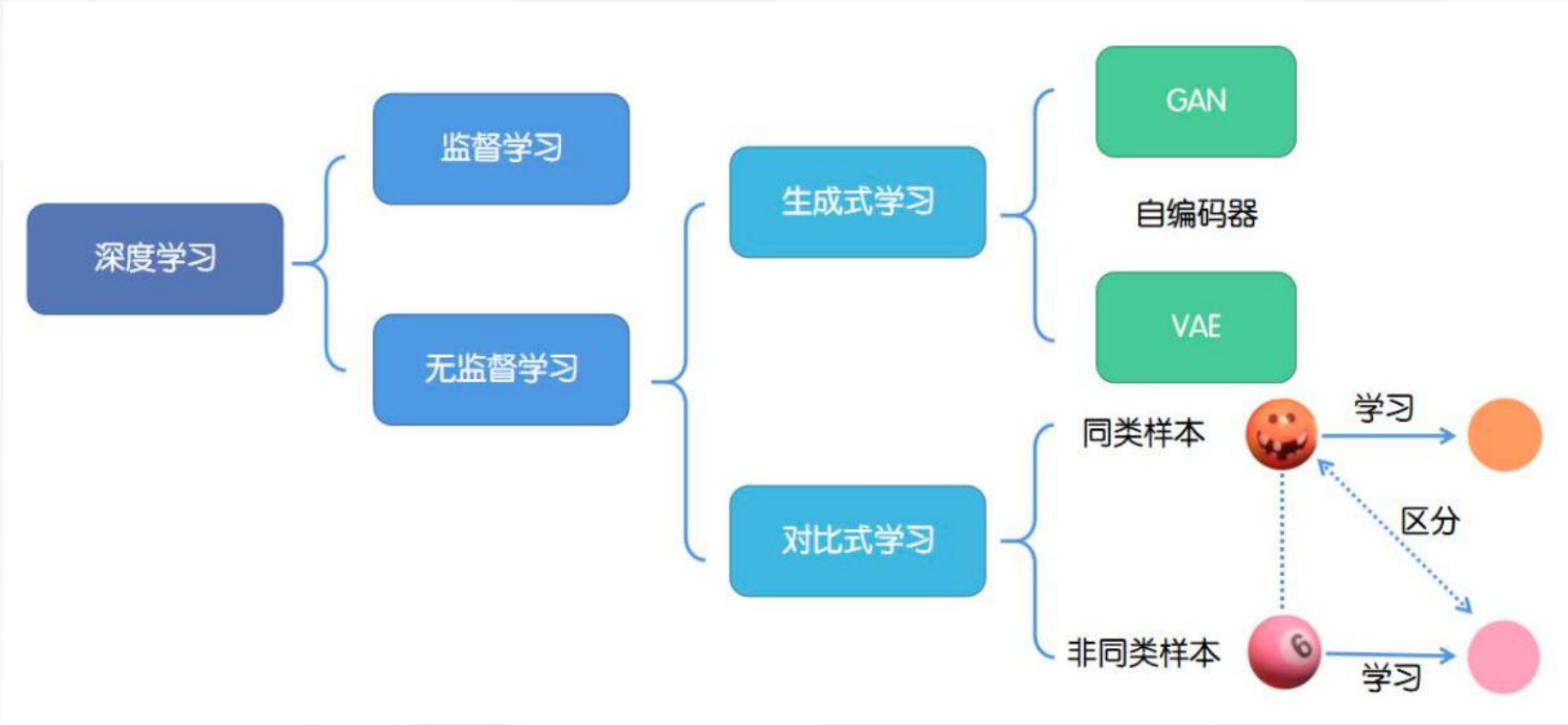
SJTU

自监督学习

徐国整

2021年11月12日

什么是自督学习？



特征：自监督学习与监督学习相对，即不需要标签。

目标：目前我关注的自监督学习，训练一个pretrain model，应用于下游任务

自监督学习范式

- Pretext tasks
  - Rotation
  - Relative patched
  - Colorization
- Generative Learning
  - Autoregressive models: PixelCNN, PixelRNN
  - Autoencoding models: Denoising autoencoder, Context encoder, Split brain autoencoder, Variational autoencoders
  - Generative Adversarial Models
- Contrastive Learning（MI、infoNCE）
  - Contrastive Predictive Coding(CPC)
  - PIRL
  - Moco、Moco v2
  - SimCLR
  - SwAV
  - BYOL

Contrastive Learning进程

- 第一阶段：概念提出
  - LeCun(2006)年提出对比损失，对比两个数据的相似性（作为降维方法）
- 第二阶段：从Pair到N-Pair
  - Contrastive MultiView Coding、Augmented Multiscale Deep InfoMax
  - Triplet Loss（2015）

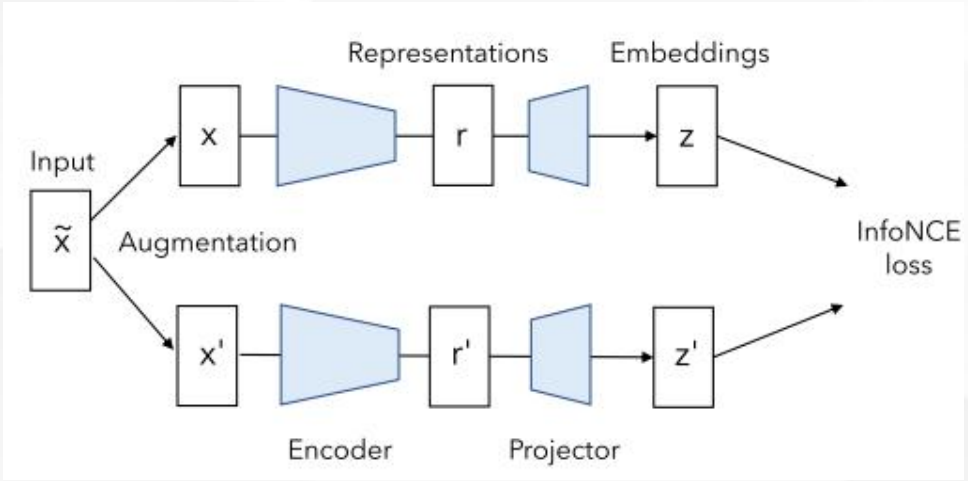
$$L = max\{d(x, x^+) - d(x, x^-) + \alpha, 0\}$$

- 第三阶段：Contrastive Representation Learning

$$\mathbb{E}[-\log \frac{f(x, x^+)}{f(x, x^+) + \sum_{k=1}^{K} f(x, x_k^-)}]$$

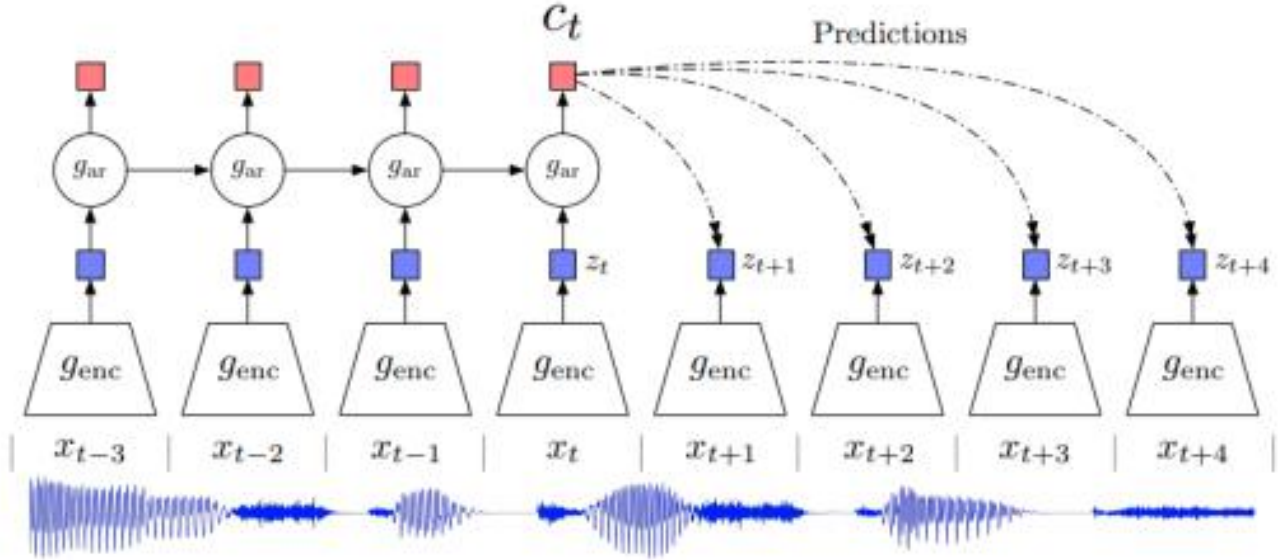
对比学习范式的研究内容

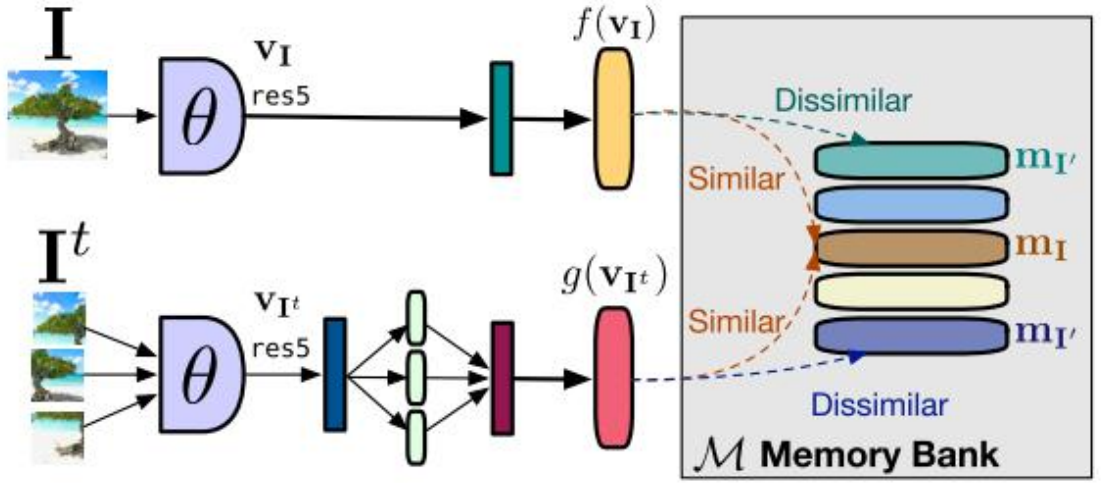
- data
  - Object-centric
  - Scene-centric
- Augmentation
  - Strong
  - Weak
- Backbone
  - CNN
  - Transformer
- FC-projector/predictor
  - One or two
  - Yes or no
- Loss
  - Positive(define)
  - Negative(need or not)
  - Tau(why)
- Training
  - Stop gradient(moment encoder)
  - Asymmetric structure
  - Batchsize, epoch
- Optimize purpose
  - Uniform and Align
  - Invariance and Variance



$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

CPC/PIPL





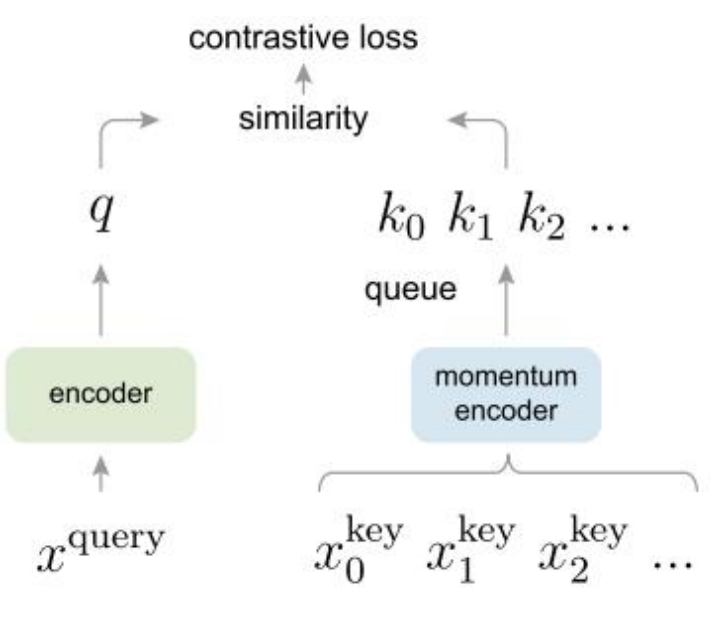
CPC

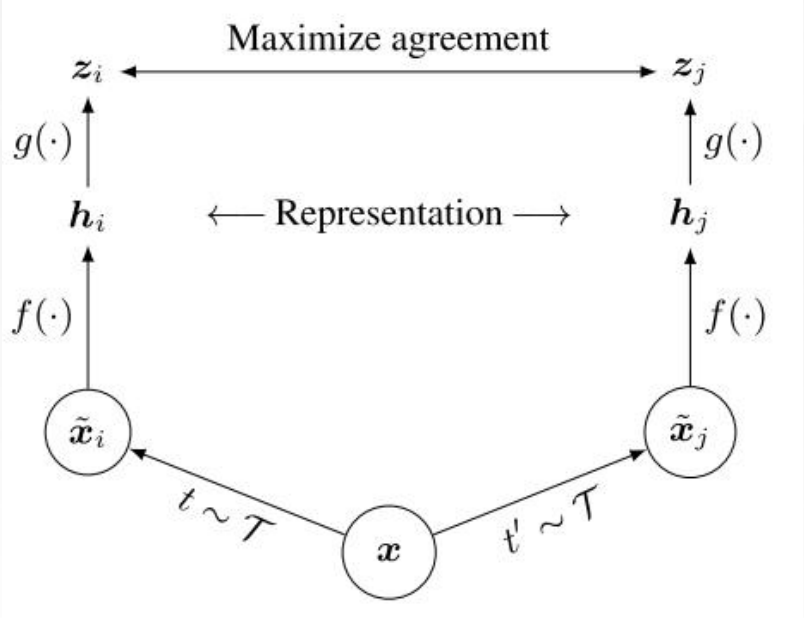
$$L_N = -\mathbb{E}_X[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}] = -\mathbb{E}_X[\log \frac{\exp(z_{t+k}^T (W_k c_t))}{\sum \exp(z_j^T (W_k c_t))}]$$

PIPL

- Memory bank
- Moment negative sample pairs

Moco/SimCLR

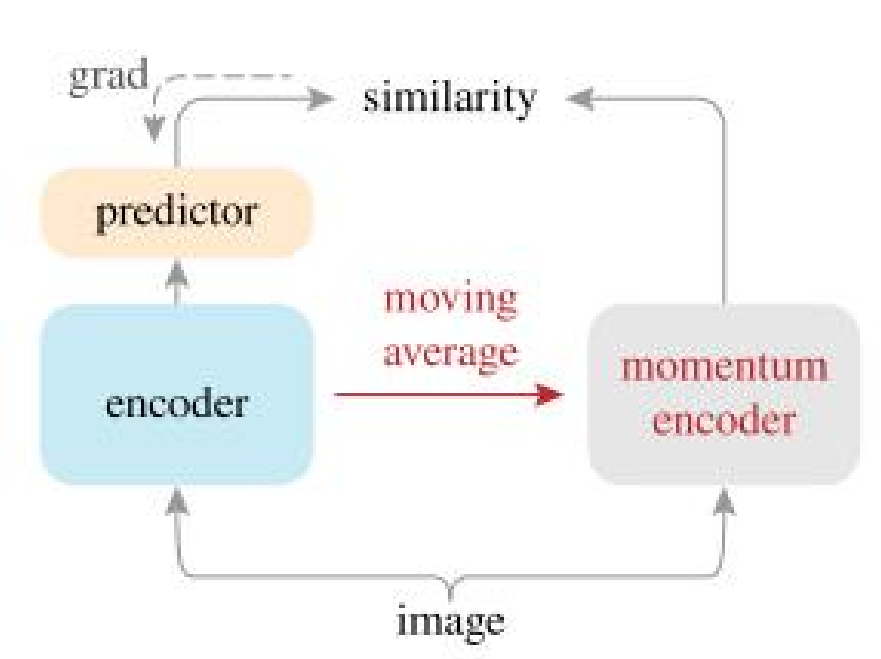


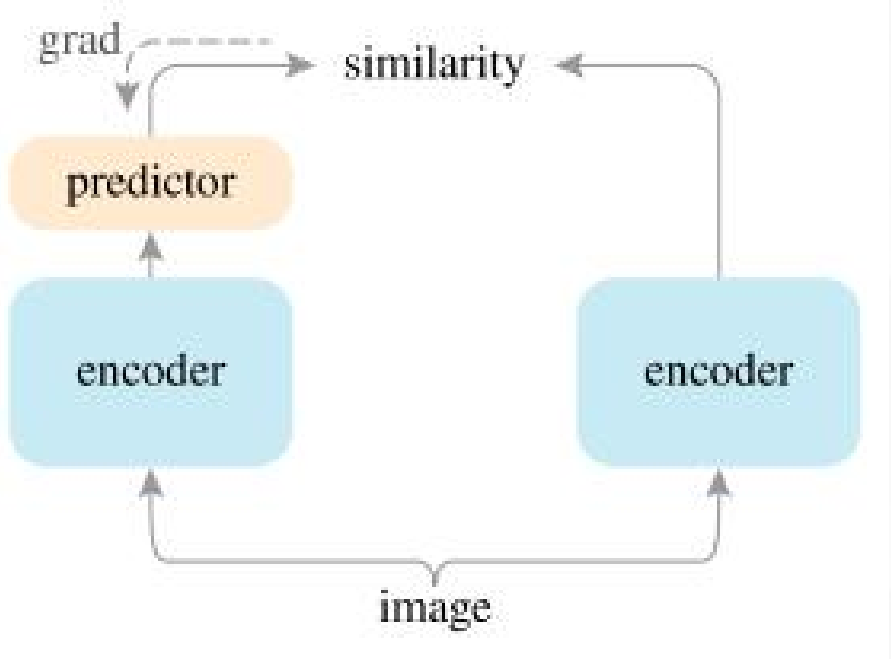


- Moco
  - Memory bank
  - Moment encoder
  - Negative sample pairs
  - Symmetric structure

- SimCLR
  - Large batches
  - Negative sample pairs
  - Symmetric structure

BYOL/SimSiam



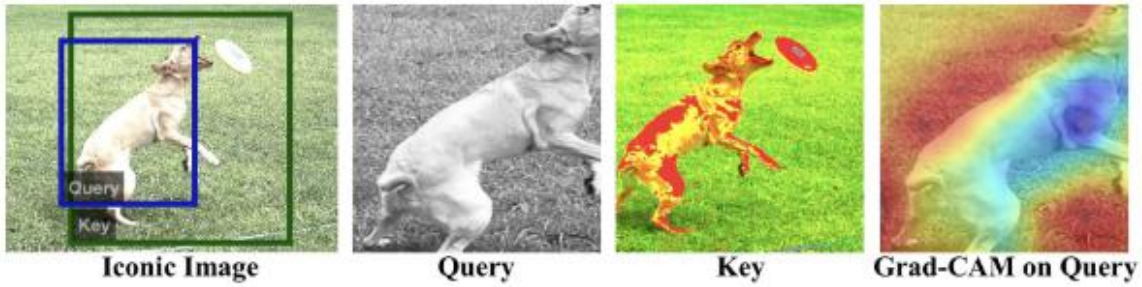


- BYOL
  - Momentum encoder
  - Asymmetric structure
- SimSiam
  - Stop gradient
  - Asymmetric structure

Data

- Multi-crop
- Constrained crop

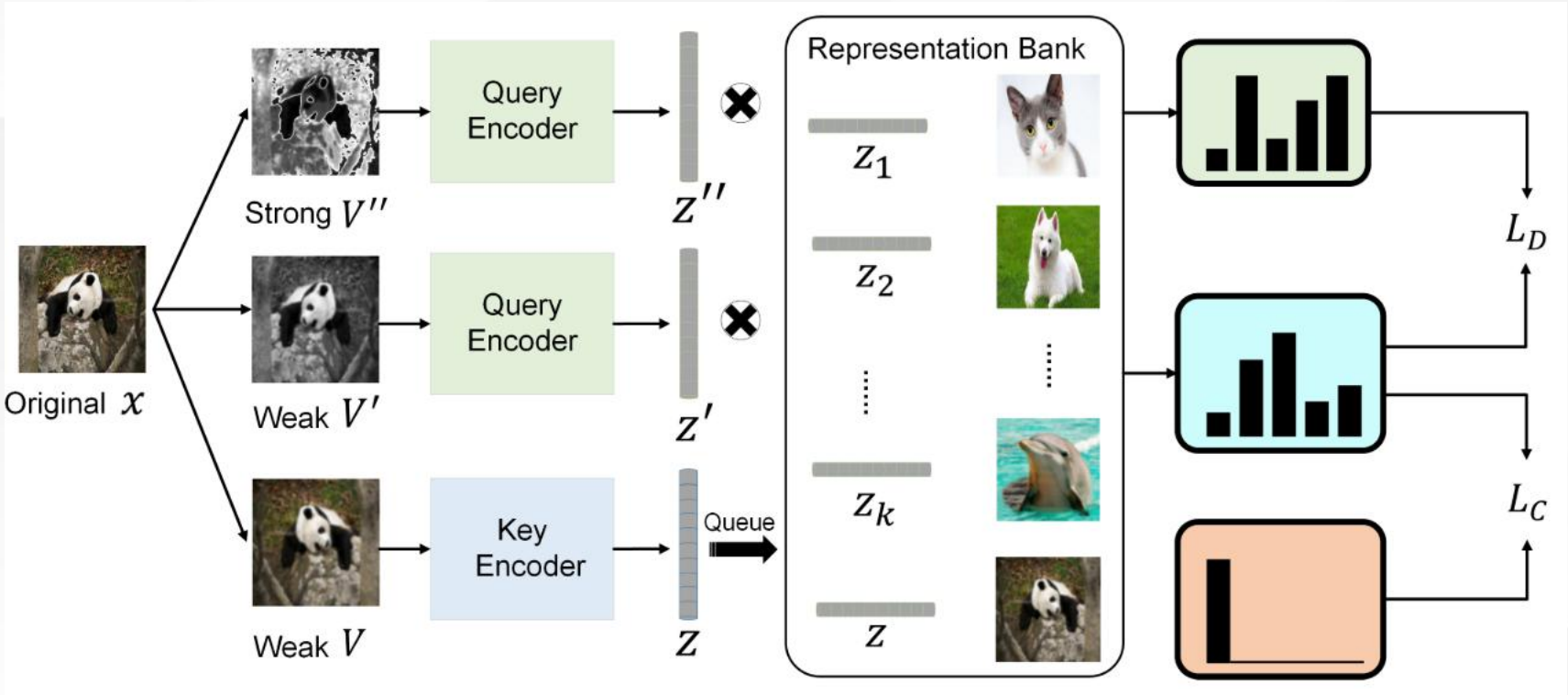
- CAM-based



(a) Poor visual grounding ability

Augmentation

- 利用更强的增强（在SimClr里面已经做了大量的实验，太弱和太强的增强都不好）：



$$\mathcal{L}_C = \mathbb{E}_{i \in B}[-q(z_i|z_i') \log p(z_i|z_i') - \sum_{k=1}^{K} q(z_k|z_i') \log p(z_k|z_i')]$$

$$\mathcal{L}_D = \mathbb{E}_{i \in B}[-p(z_i|z_i') \log p(z_i|z_i'') - \sum_{k=1}^{K} p(z_k|z_i') \log p(z_k|z_i'')]$$

Loss-梯度求导

- 解释对比损失为什么需要大量负样本的原因，如果才能不需要

$$L_i^{(k)} = -\log \frac{\exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \rangle / \tau)}{\exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \rangle / \tau) + \sum_{l \in \{1,2\}, j \in [\![1,N]\!], j \neq i} \exp(\langle \mathbf{z}_i^{(k)}, \mathbf{z}_j^{(l)} \rangle / \tau)}$$

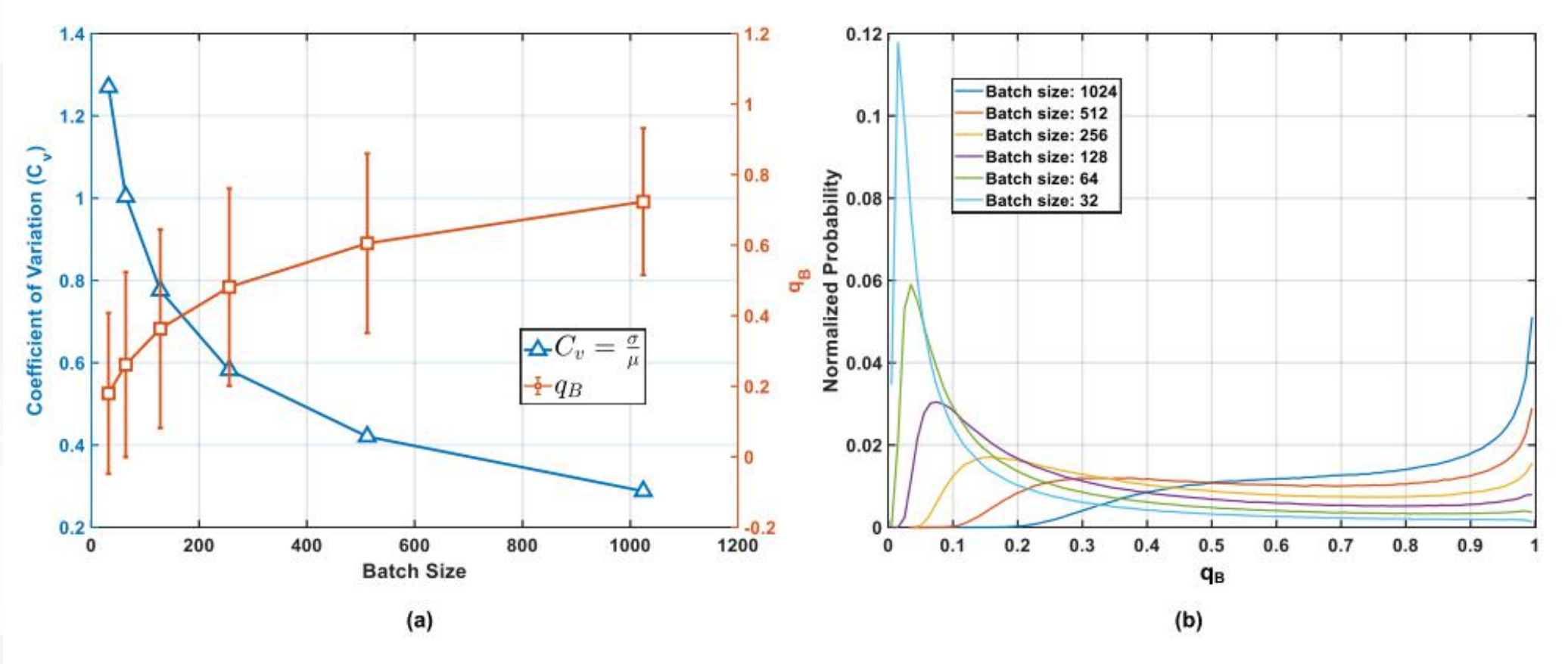
$$\begin{cases} -\nabla_{\mathbf{z}_i^{(1)}} L_i^{(1)} = \frac{q_{B,i}^{(1)}}{\tau} \left[\mathbf{z}_i^{(2)} - \sum_{l \in \{1,2\}, j \in [\![1,N]\!], j \neq i} \frac{\exp \langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(l)} \rangle / \tau}{\sum_{q \in \{1,2\}, j \in [\![1,N]\!], j \neq i} \exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(q)} \rangle / \tau)} \cdot \mathbf{z}_j^{(l)} \right] \\ -\nabla_{\mathbf{z}_i^{(2)}} L_i^{(1)} = \frac{q_{B,i}^{(1)}}{\tau} \cdot \mathbf{z}_i^{(1)} \\ -\nabla_{\mathbf{z}_j^{(l)}} L_i^{(1)} = -\frac{q_{B,i}^{(1)}}{\tau} \frac{\exp \langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(l)} \rangle / \tau}{\sum_{q \in \{1,2\}, j \in [\![1,N]\!], j \neq i} \exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(q)} \rangle / \tau)} \cdot \mathbf{z}_i^{(1)} \end{cases}$$

- 解决方案

$$L_{DCW,i}^{(k)} = -w(\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)})(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \rangle / \tau) + \log \sum_{l \in \{1,2\}, j \in [\![1,N]\!], j \neq i} \exp(\langle \mathbf{z}_i^{(k)}, \mathbf{z}_j^{(l)} \rangle / \tau)$$

Loss-梯度求导

- Bs小的时候，qB会有更大的概率接近0，这时候就不优化了



Loss-tau

- 解释对比损失中tau到底有什么作用

$$P_{i,j} = \frac{\exp(s_{i,j}/\tau)}{\sum_{k\neq i}\exp(s_{i,k}/\tau) + \exp(s_{i,i}/\tau)}$$

$$\mathcal{L}_{simple}(x_i) = -s_{i,i} + \lambda \sum_{i\neq j} s_{i,j}$$

$$\lim_{\tau\to 0^+} -\log\left[\frac{\exp(s_{i,i}/\tau)}{\sum_{k\neq i}\exp(s_{i,k}/\tau) + \exp(s_{i,i}/\tau)}\right]$$

$$= \lim_{\tau\to 0^+} +\log\left[1 + \sum_{k\neq i}\exp((s_{i,k} - s_{i,i})/\tau)\right]$$

$$= \lim_{\tau\to 0^+} +\log\left[1 + \sum_{s_{i,k}\geqslant s_{i,i}}^{k}\exp((s_{i,k} - s_{i,i})/\tau)\right]$$

$$= \lim_{\tau\to 0^+} \frac{1}{\tau} max[s_{max} - s_{i,i}, 0]$$

$$\lim_{\tau\to +\infty} -\log\left[\frac{\exp(s_{i,i}/\tau)}{\sum_{k\neq i}\exp(s_{i,k}/\tau) + \exp(s_{i,i}/\tau)}\right]$$

$$= \lim_{\tau\to +\infty} -\frac{1}{\tau}s_{i,i} + \log\sum_{k}\exp(s_{i,k}/\tau)$$

$$= \lim_{\tau\to +\infty} -\frac{1}{\tau}s_{i,i} + \frac{1}{N}\sum_{k}\exp(s_{i,k}/\tau) - 1 + \log N$$

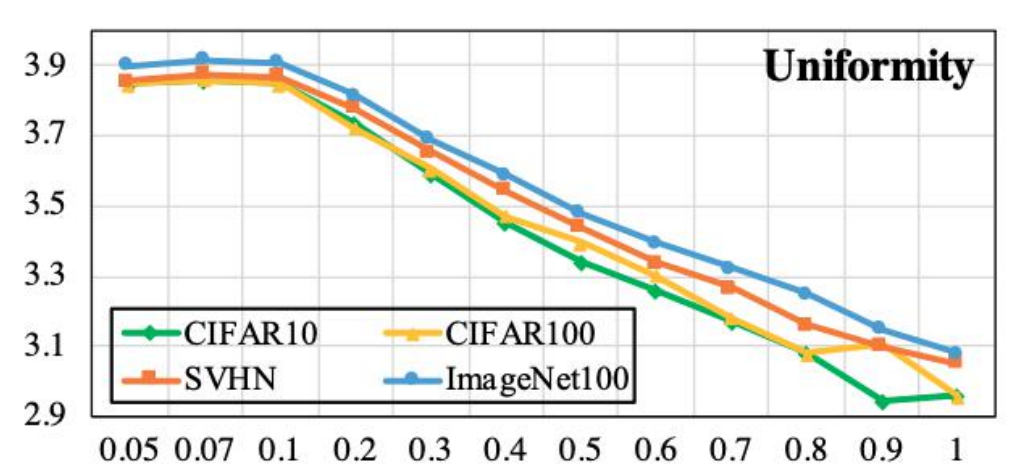
$$= \lim_{\tau\to +\infty} -\frac{N-1}{N\tau}s_{i,i} + \frac{1}{N\tau}\sum_{k\neq i} s_{i,k} + \log N$$

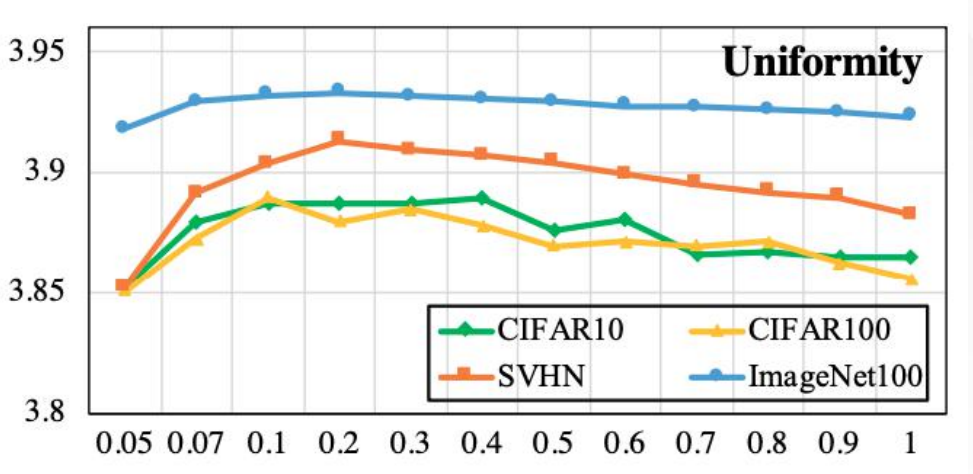
- 解决方案

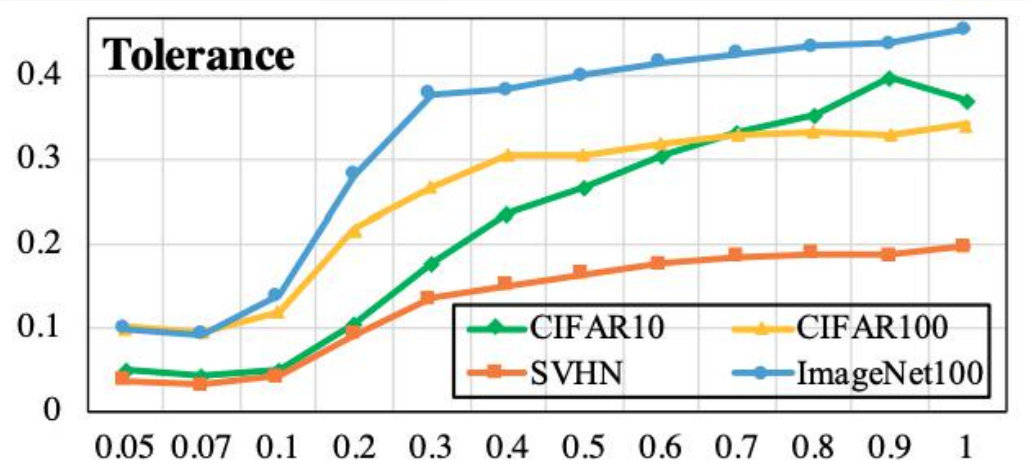
$$\mathcal{L}_{\text{hard}}(x_i) = -\log\frac{\exp(s_{i,i}/\tau)}{\sum_{s_{i,k}\geqslant s_{\alpha}^{(i)}}\exp(s_{i,k}/\tau) + \exp(s_{i,i}/\tau)}$$

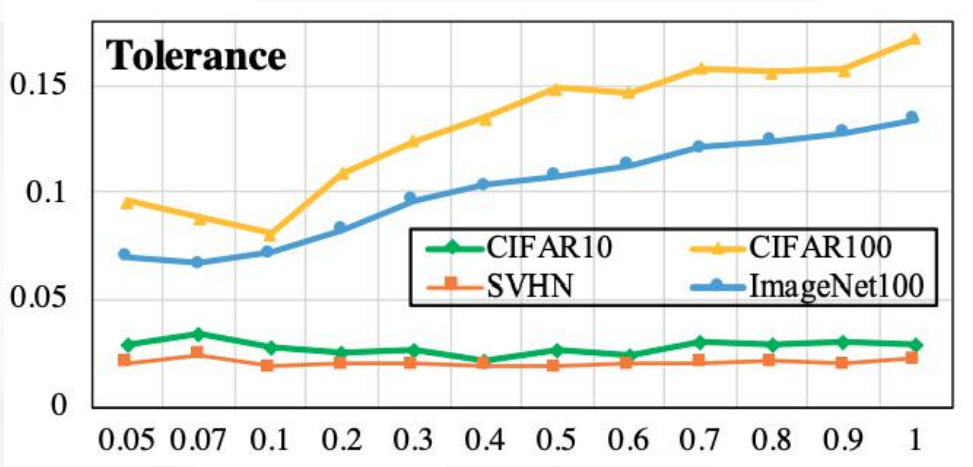
Loss-tau

- 解释对比损失中tau到底有什么作用（分析性文章，性能基本上没有提升）





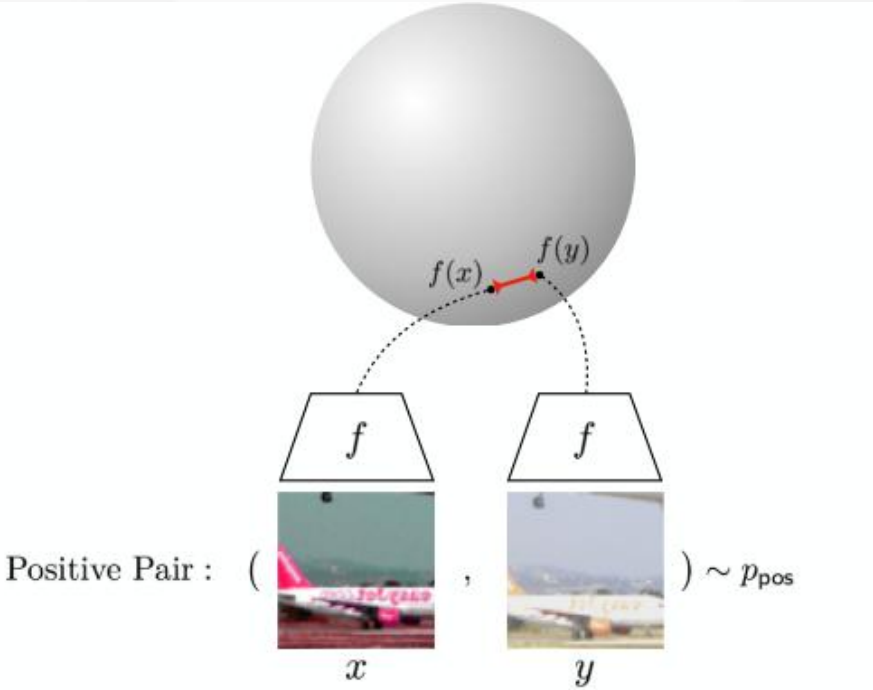




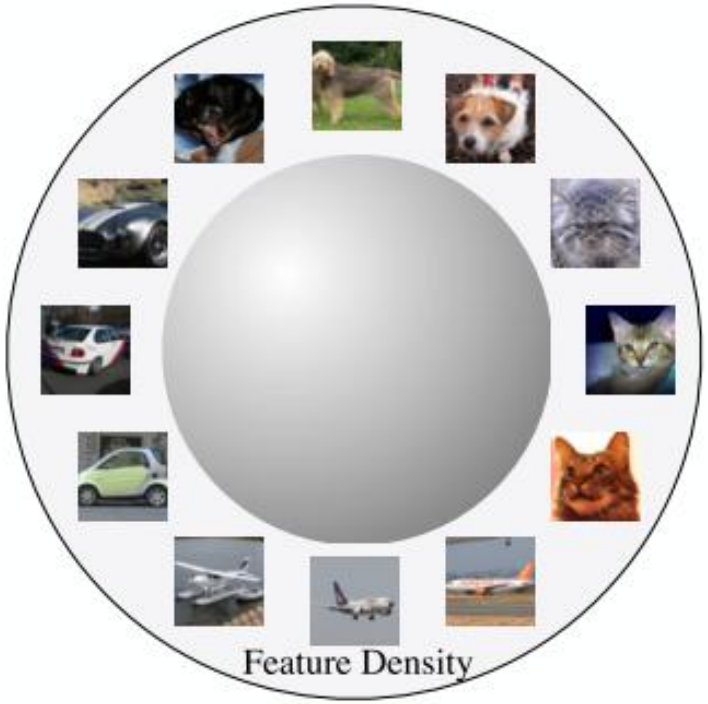
$$\mathcal{L}_{\text{uniformity}}(f;t) = \log \mathop{\mathbb{E}}_{x,y\sim p_{data}} \left[e^{-t\|f(x)-f(y)\|_2^2} \right]$$

$$T = \mathop{\mathbb{E}}_{x,y\sim p_{data}} \left[(f(x)^T f(y)) \cdot I_{l(x)=l(y)} \right]$$

Loss-对齐和均匀性



Alignment: Similar samples have similar features.
(Figure inspired by Tian et al. (2019).)



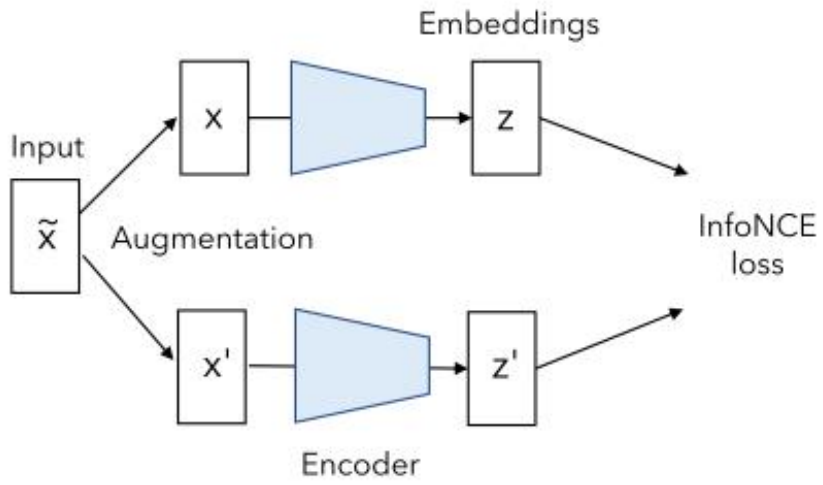
Uniformity: Preserve maximal information.

$$\mathcal{L}_{\text{align}}(f;\alpha) \triangleq \mathbb{E}_{(x,y)\sim p_{\text{pos}}}\left[\|f(x)-f(y)\|_2^{\alpha}\right]$$

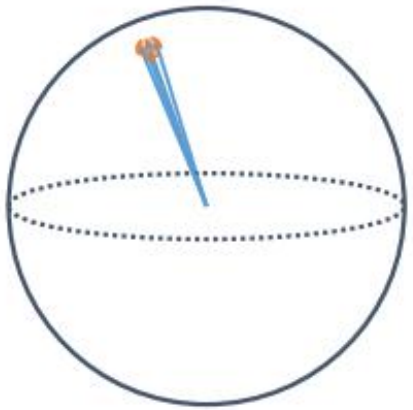
$$\mathcal{L}_{\text{uniform}}(f;t) \triangleq \log \mathbb{E}_{x,y \overset{\text{i.i.d.}}{\sim} p_{\text{data}}}\left[e^{-t\|f(x)-f(y)\|_2^2}\right]$$

Projector-dimensional collapse?

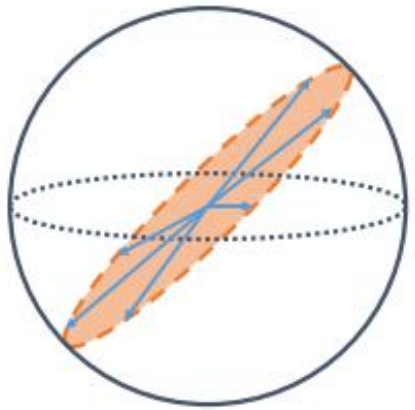
We showed that there are two mechanisms causing the dimensional collapse in contrastive learning: (1) strong augmentation along feature dimensions (2) implicit regularization driving models toward low-rank solutions.



(a) embedding space



(b) complete collapse



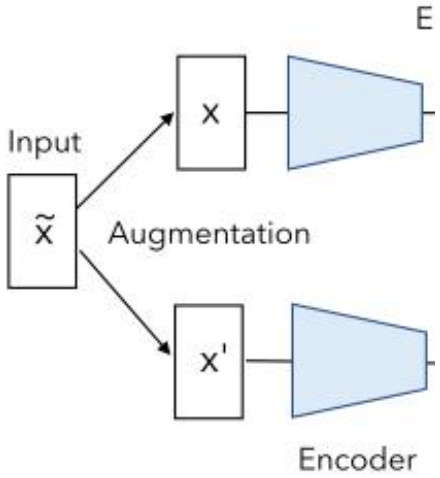
(c) dimensional collapse

$$C = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T$$

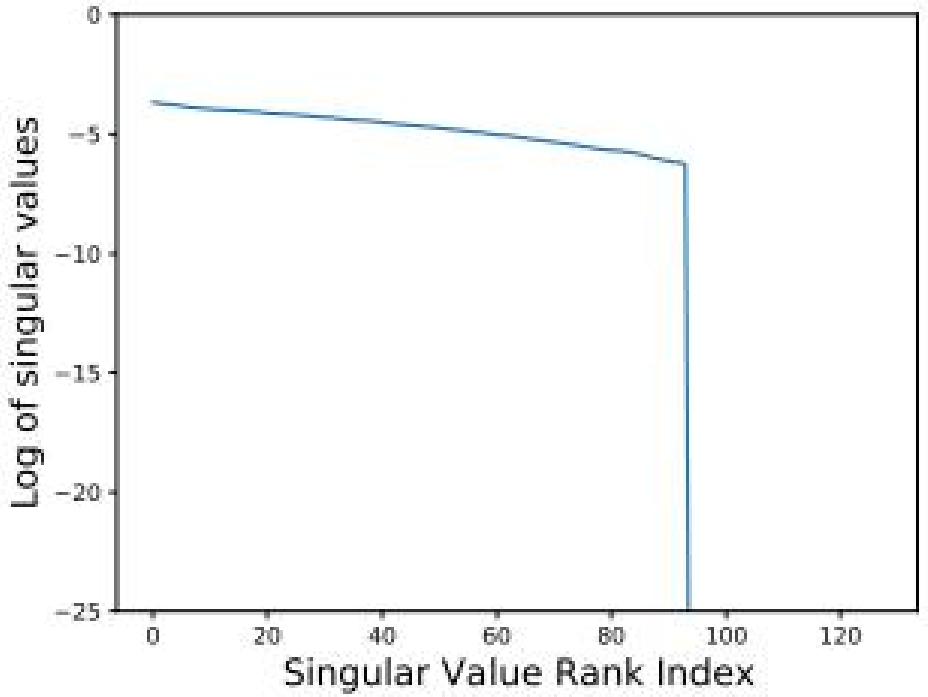
$$C = USV^T, S = diag(\sigma^k)$$

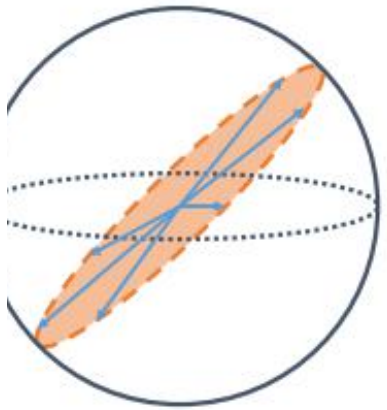
Projector-dimensional collapse?

We showed that there are two mechanisms causing the dimensional collapse in contrastive learning: (1) strong a[illegible] it regularization driving models toward lo[illegible]



(a) embeddi[illegible]



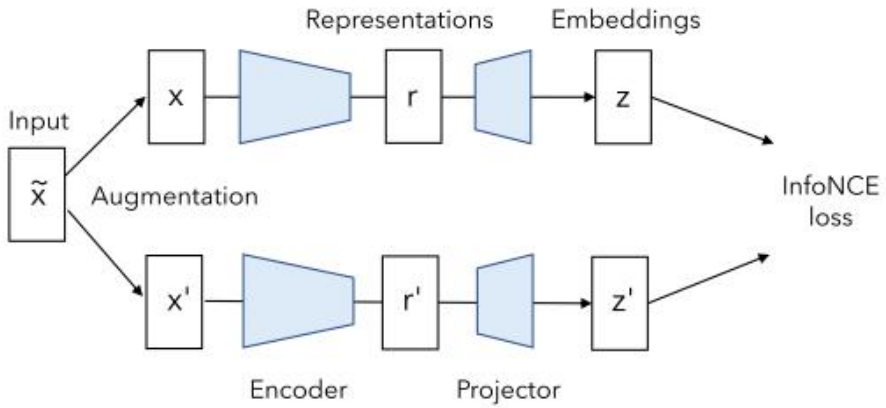


dimensional collapse

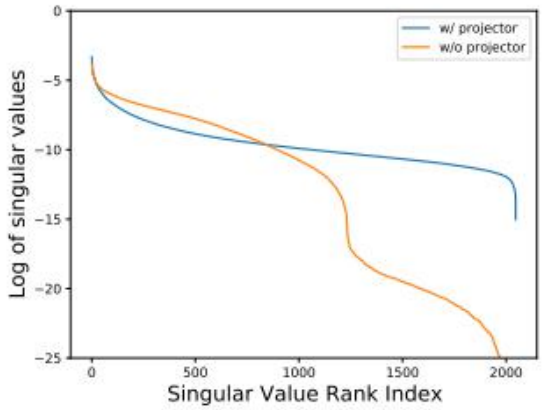
$$C = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T$$

$$C = USV^T, S = diag(\sigma^k)$$

Projector-dimensional collapse?



(a) representation and embedding



(b) Representation space spectrum

Figure 7: (a) Definition of representation and the embedding space; (b) Singular value spectrums of the representation space of pretrained contrastive learning models (pretrained with or without a projector). The representation vectors are the output from the ResNet50 encoder and directly used for downstream tasks. Each representation vector has a dimension of 2048. Without a projector, SimCLR suffers from dimensional collapse in the representation space.

1. The gradient will drive the projector weight matrix aligned with the last layer of the encoder backbone. We suspect that such alignment effect $V_2^T U_1 \to I$ only requires one of V_2 and U_1 to evolve. Therefore, the projector weight matrix only needs to be **diagonal**.
2. The projector only applies a gradient to a subspace to the representations. Therefore, the projector weight matrix only needs to be **low-rank**.

Projector-dimensional collapse?

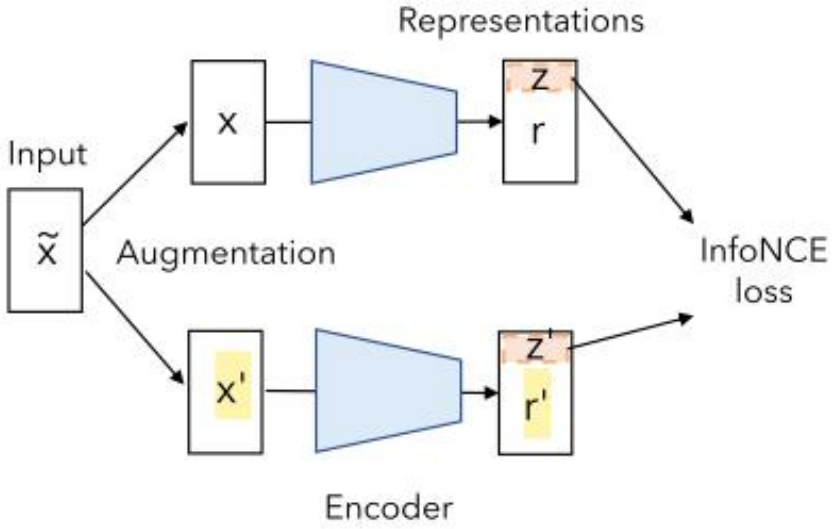
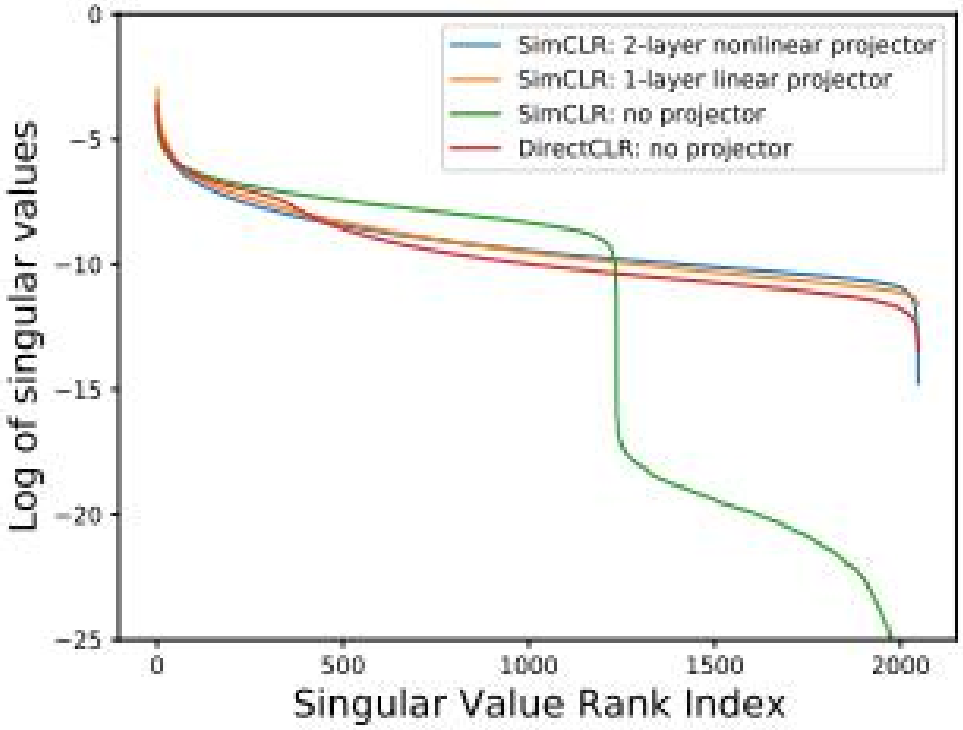
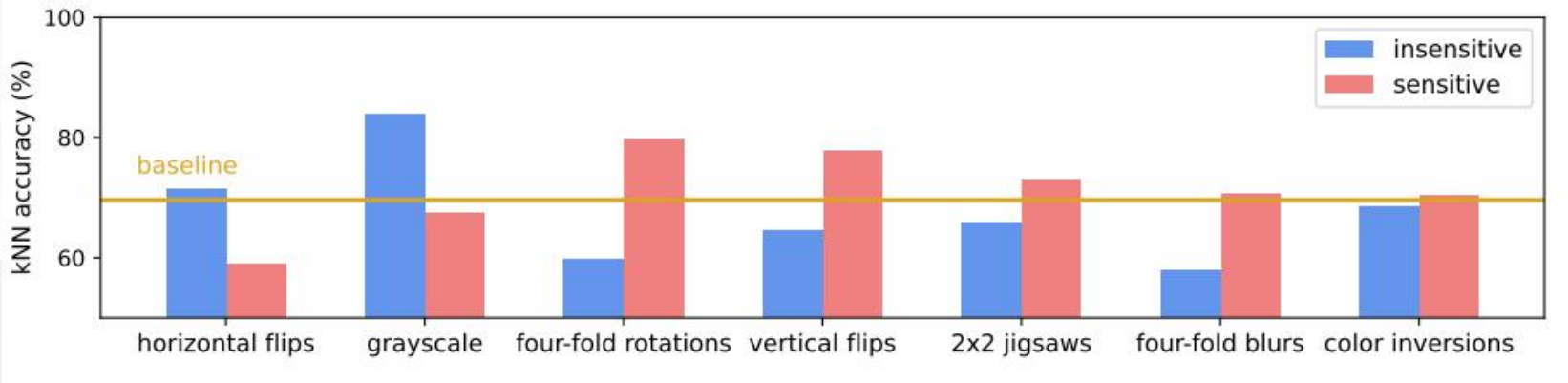
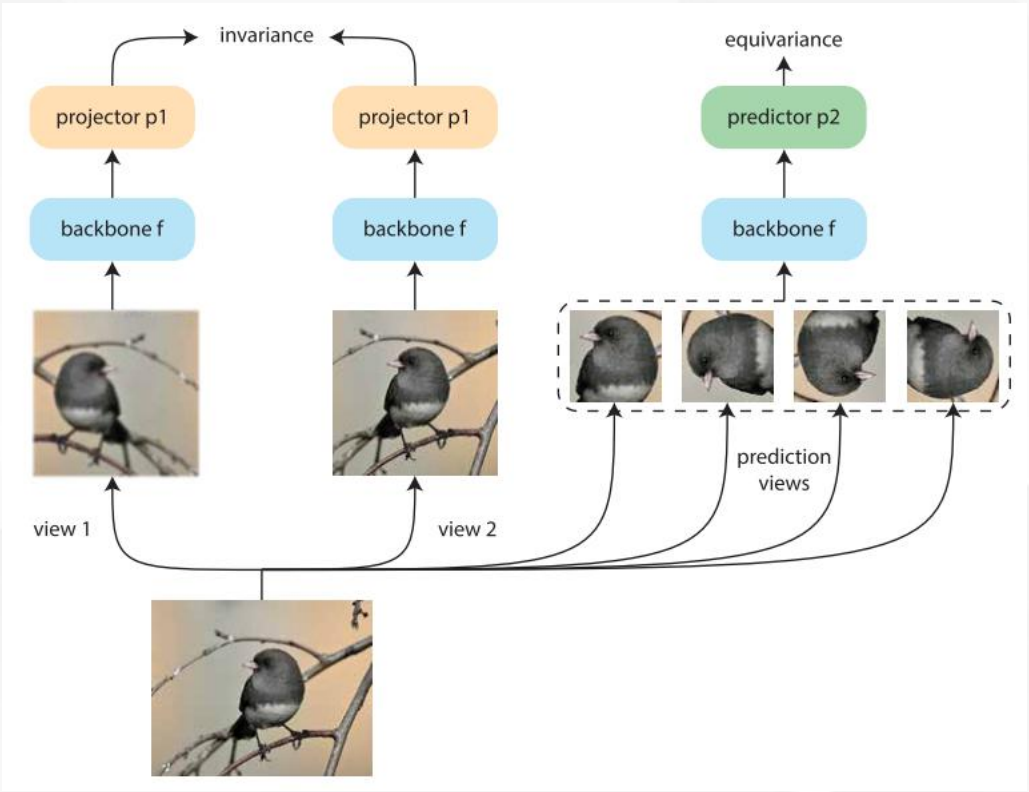


Figure 8: *DirectCLR*: no trainable projector, simply apply InfoNCE loss on the a fixed sub-vector of the representations



不变性和等变性


100
80
60
kNN accuracy (%)
baseline
insensitive
sensitive
horizontal flips
grayscale
four-fold rotations
vertical flips
2x2 jigsaws
four-fold blurs
color inversions


invariance
equivariance
projector p1
projector p1
predictor p2
backbone f
backbone f
backbone f
prediction
views
view 1
view 2

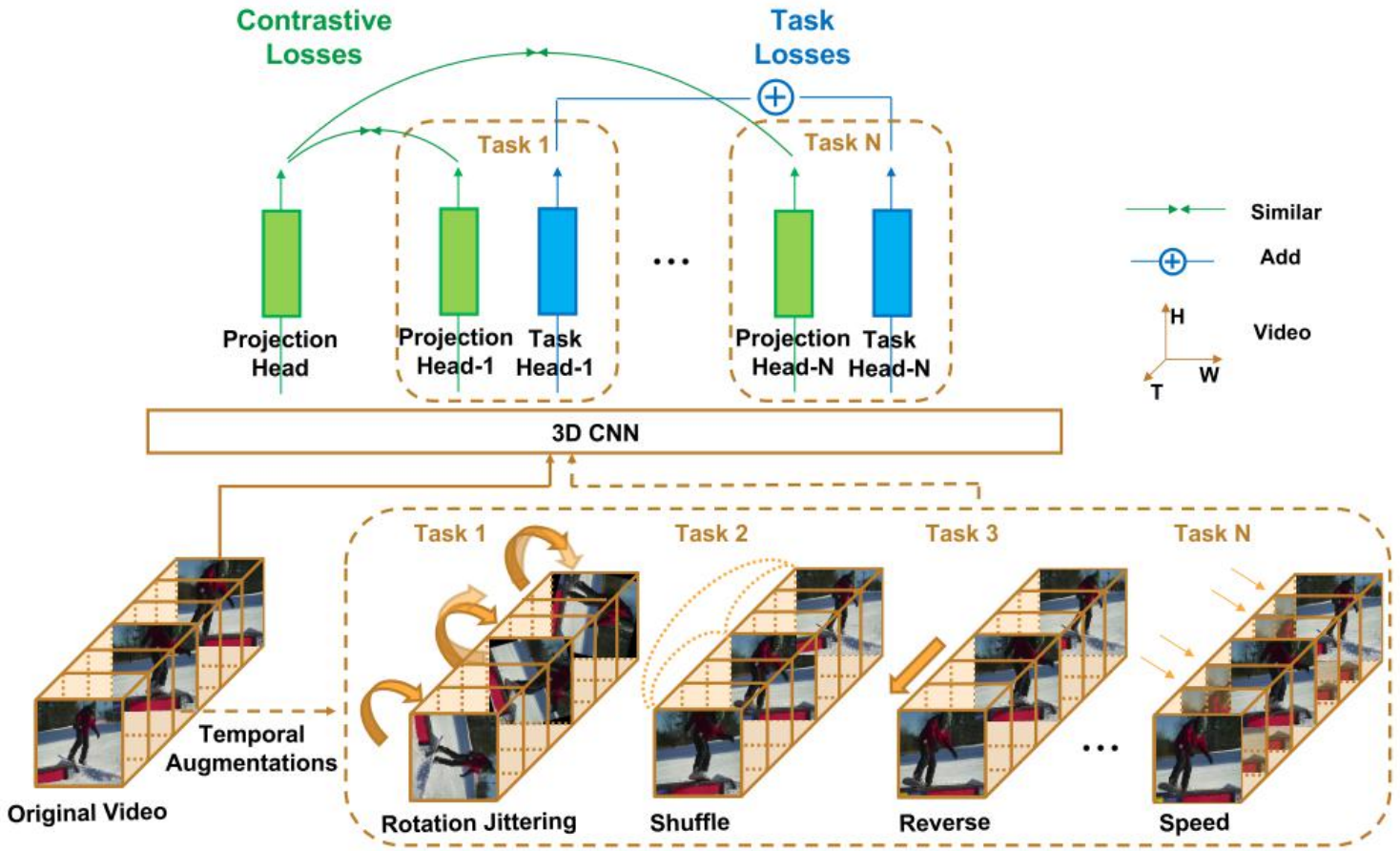
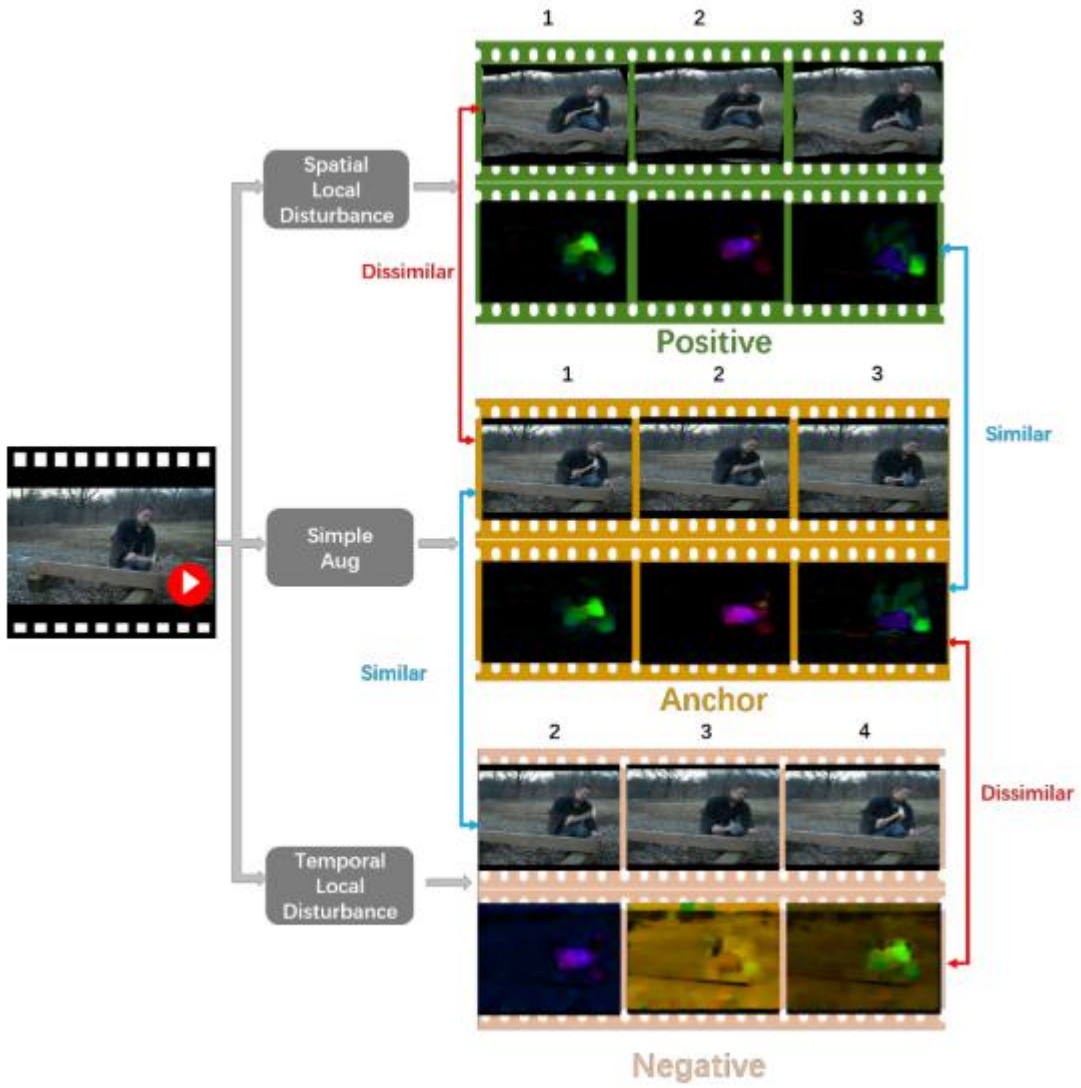
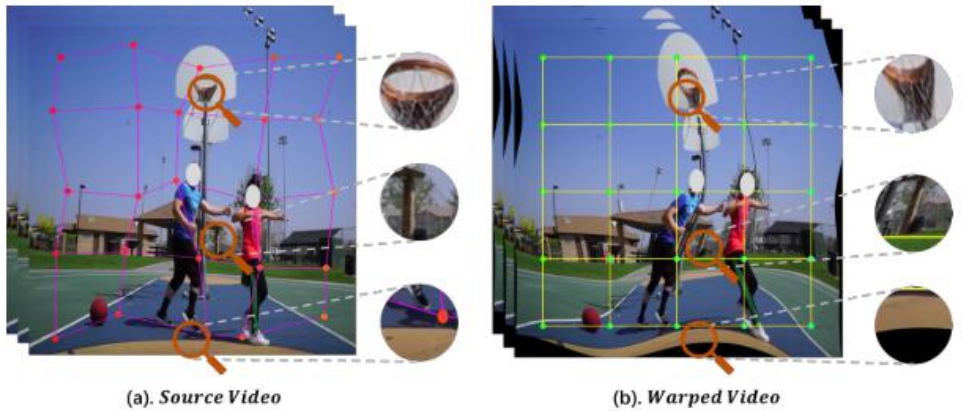


Figure 2. **Overview of the proposed temporal-aware contrastive self-supervised learning framework (TaCo).** TaCo mainly comprises three modules: temporal augmentation module, contrastive learning module, and temporal pretext task module. For different temporal augmentations, we apply different projection heads and task heads. The features extracted from projection head of original video sequence and augmented sequence are considered as positive sample pairs, and the remaining ones are simply regarded as negative sample pairs. The contrastive loss is computed as the summation of losses over all pairs.

Video-BE



Spatial Local Disturbance



(a). *Source Video* (b). *Warped Video*

Figure 3: Illustration of the **Spatial Warping**, which randomly warps spatial regions in each epoch. Though the local statistics is broken, the global statistics is maintained.

Optical-flow Scaling.

$$\frac{\partial I}{\partial x}V_x + \frac{\partial I}{\partial y}V_y + \frac{\partial I}{\partial t} = 0,$$

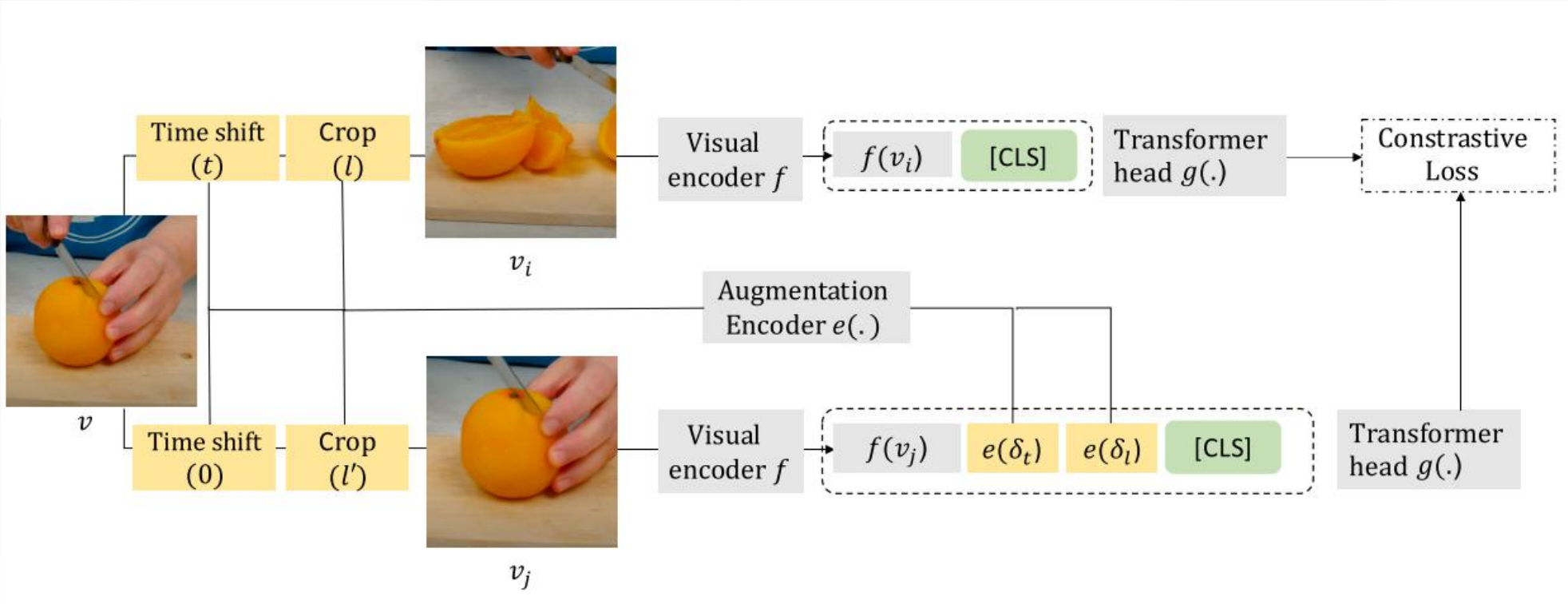
$$I(x, y, t+1) = I(x + V_x, y + V_y, t)$$

$$\hat{I}(x, y, t+1) = I(x + \phi(t)V_x, y + \phi(t)V_y, t)$$

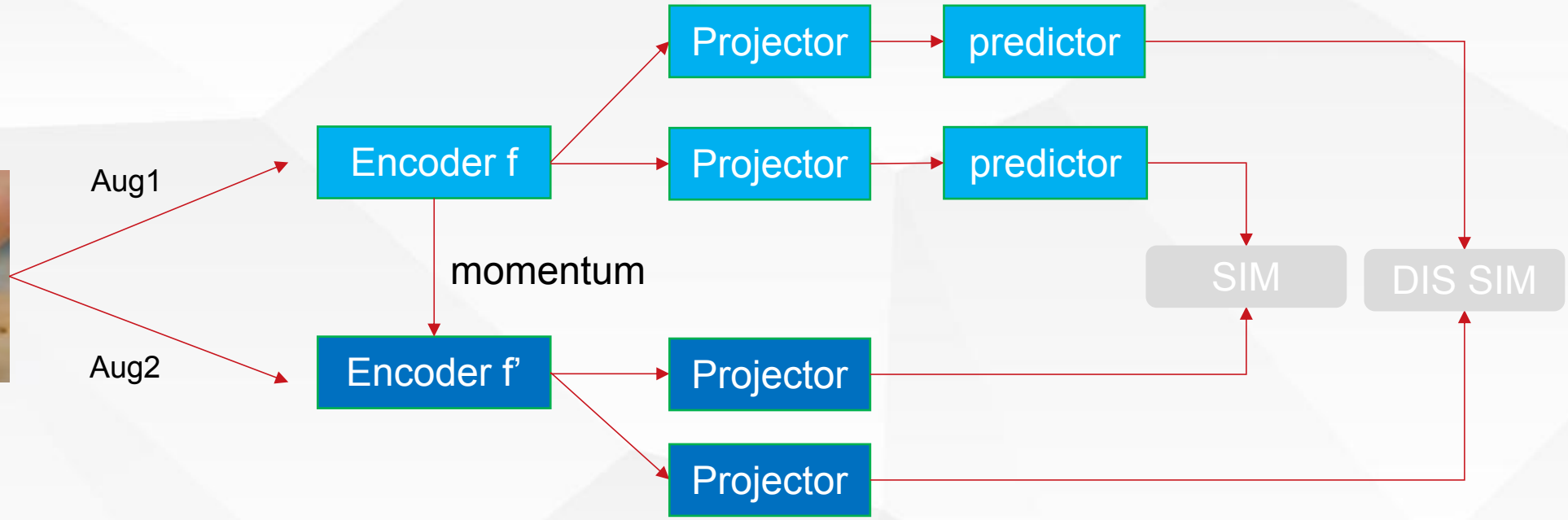
Temporal Shift.

$$\hat{x}_i = x_{i+\tau}, i \in 1, 2...T$$

Video-不变性


Time shift
(t)
Crop
(l)
v_i
Visual
encoder f
$f(v_i)$
[CLS]
Transformer
head $g(.)$
Constrastive
Loss
v
Augmentation
Encoder $e(.)$
Time shift
(0)
Crop
(l')
v_j
Visual
encoder f
$f(v_j)$
$e(\delta_t)$
$e(\delta_l)$
[CLS]
Transformer
head $g(.)$

我的工作



gate：　$f(\delta x) = e^{-\frac{|\delta x|}{T}}$

一致性：　$x = x + f(\delta x) * x$

非一致性：　$x = x + (1 - f(\delta x)) * x$

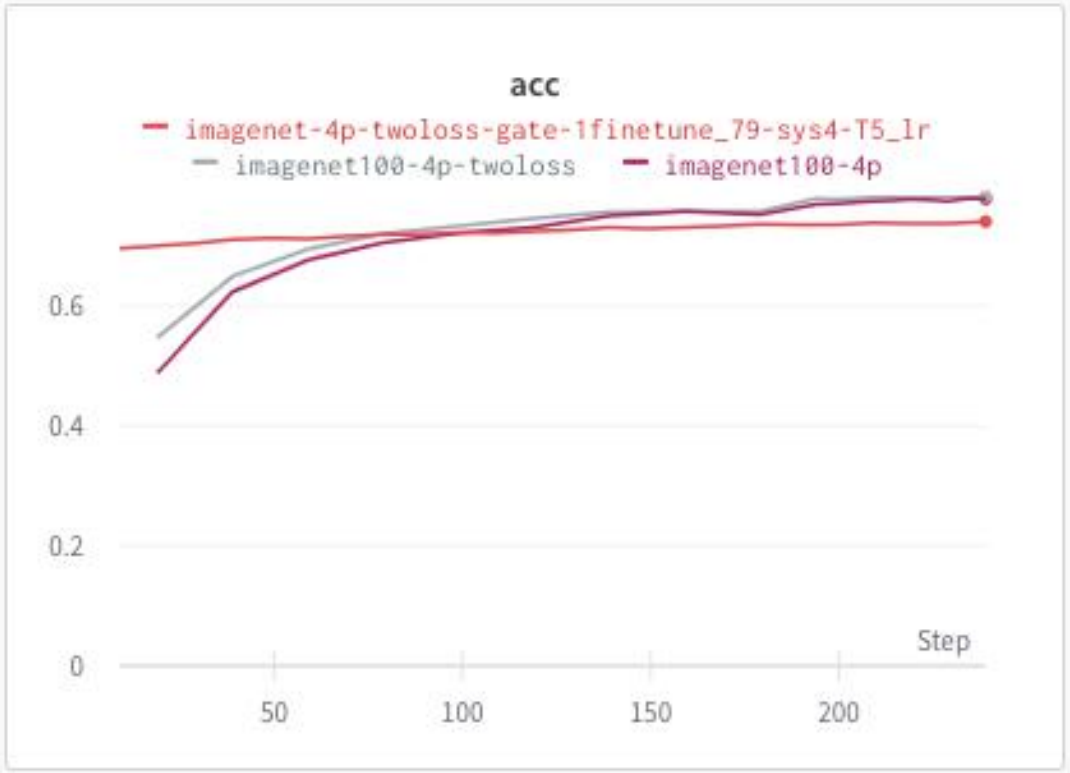
我的工作

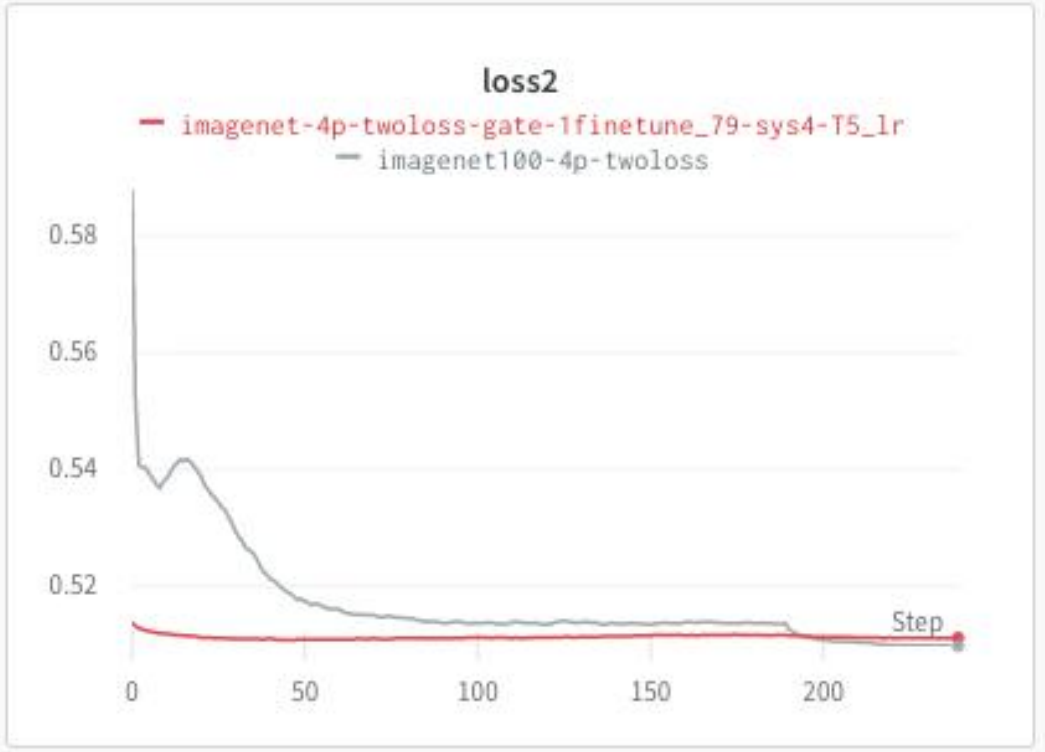
数据集	方法	epoch	backbone	bs	acc	acc_5	acc_knn	loss	loss1	loss2
stl(96)	byol+loss1	2000	resnet18	4096	0.9174	0.996	0.89	0.246		
stl(96)	byol+two-loss	2000	resnet18	4096	0.9108 (-0.0066)	0.996	0.878 (-0.012)	0.7781	0.2624	0.5157
cifar-100(32)	byol+loss1	1000	resnet18	4096	0.6868	0.9134	0.6164	0.3915		
cifar-100(32)	byol+two-loss	1000	resnet18	4096	0.6839 (-0.0029)	0.9093 (-0.0041)	0.6164	0.9035	0.3831	0.5203
cifar-100(32)	byol+loss2	1000	resnet18	4096	0.5749 (-0.1119)	0.8436 (-0.0698)	0.4724 (-0.144)	0.5181		
cifar-10(32)	byol+loss1	1000	resnet18	512	0.9159	0.9973	0.9013	0.2623		
cifar-10(32)	byol+two-loss	1000	resnet18	512	0.9192 (+0.0033)	0.997 (-0.0003)	0.8986 (-0.0027)	0.7711		
cifar-10(32)	byol+loss2	1000	resnet18	512	0.885 (-0.0309)	0.9966 (-0.0007)	0.8571 (-0.0442)	0.5169		

我的工作

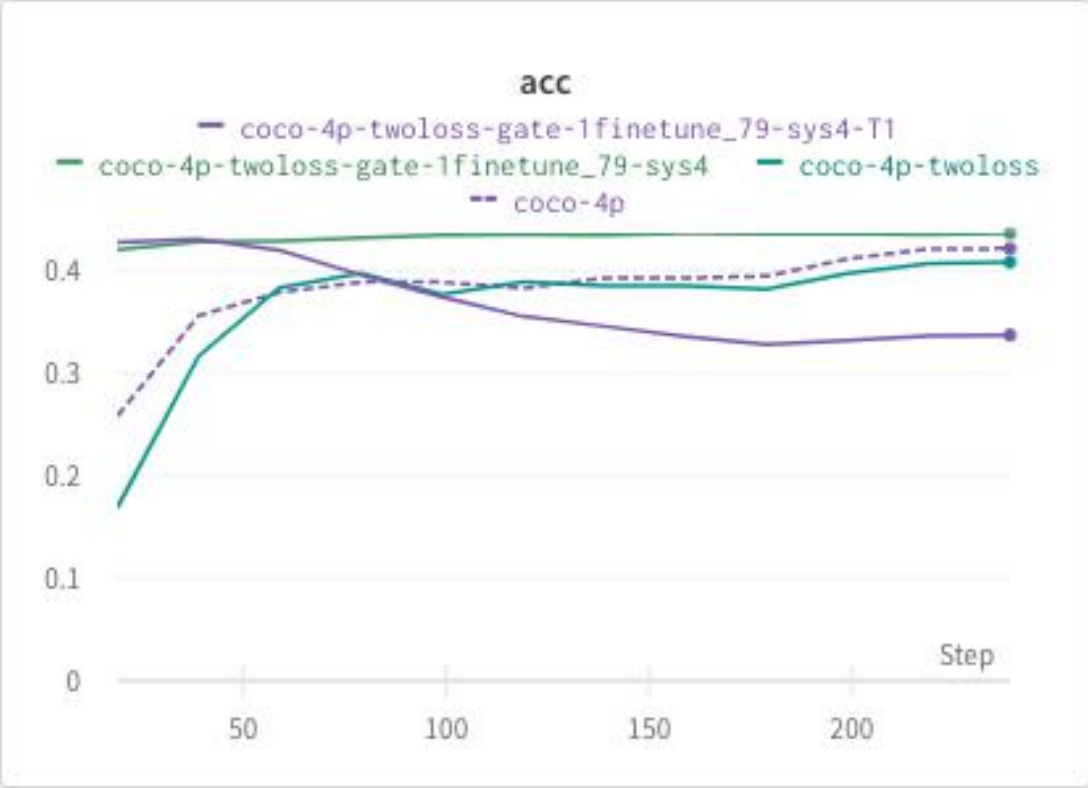
		backbone	cityscapes/pspnet			imagenet	imagenet100
数据集	pretrain method	resnet50	aAcc	mIOU	mAcc	acc	acc
coco	coco	resnet50	93.95	64.97	72.18	0.4214	
coco	coco_twoloss	resnet50	93.83	65.19	72.34	0.4082	
coco	coco_4_1finetune_79_4p	resnet50	94.36	68.11	75	0.4364	
coco	coco_4_1finetune_79_8new	resnet50	94.63	69.16	76.57	0.3371	
		resnet50					
imagenet	imagenet_0	resnet50	94.56	67.84	75.41		0.7782
imagenet	imagenet_1	resnet50	94.72	69.61	77.49		0.7828
imagenet	imagenet_4_1finetnue_79_sys4_T5_lr	resnet50	93.99	63.95	71.43		0.7404

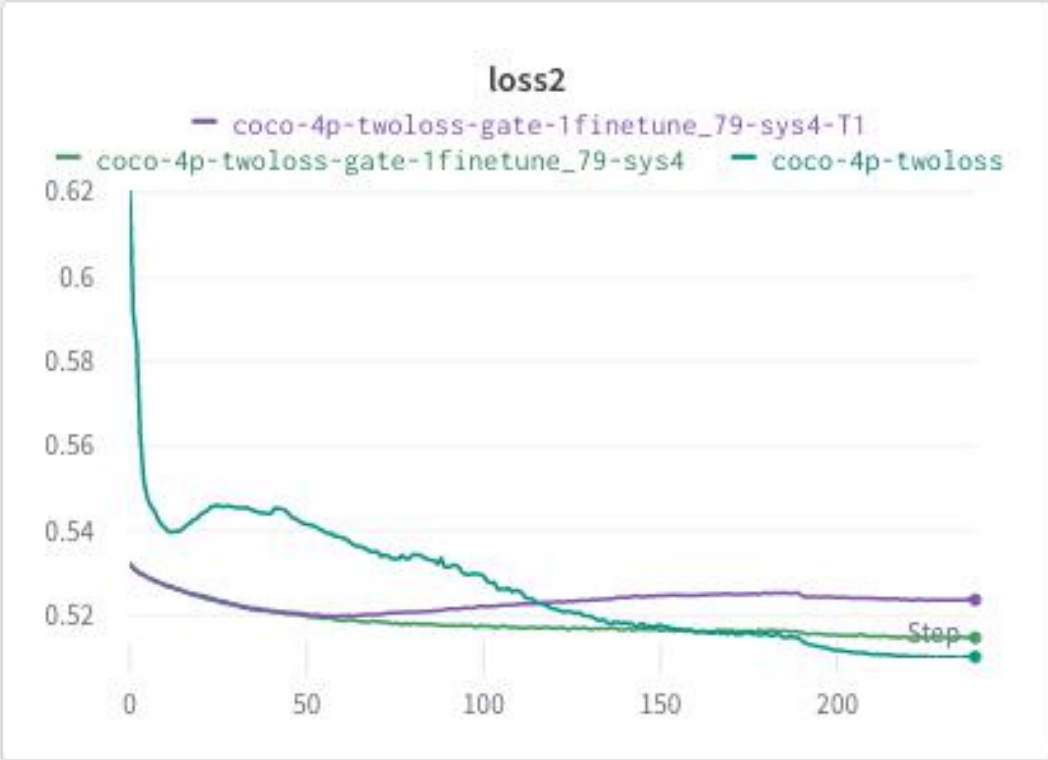
我的工作


acc
imagenet-4p-twoloss-gate-1finetune_79-sys4-T5_lr
imagenet100-4p-twoloss
imagenet100-4p
0.6
0.4
0.2
0
Step
50
100
150
200


loss2
imagenet-4p-twoloss-gate-1finetune_79-sys4-T5_lr
imagenet100-4p-twoloss
0.58
0.56
0.54
0.52
Step
0
50
100
150
200

我的工作


acc
coco-4p-twoloss-gate-1finetune_79-sys4-T1
coco-4p-twoloss-gate-1finetune_79-sys4
coco-4p-twoloss
coco-4p
0.4
0.3
0.2
0.1
0
Step
50
100
150
200


loss2
coco-4p-twoloss-gate-1finetune_79-sys4-T1
coco-4p-twoloss-gate-1finetune_79-sys4
coco-4p-twoloss
0.62
0.6
0.58
0.56
0.54
0.52
Step
0
50
100
150
200

开放问题

- 表征学习的align和uniform之间的balance问题
- Classification and Segmentation等任务之间的统一问题
- 为什么现在的网络能work的理论解释
- 在视频中，则是重点解决时间维度信息的问题

上海交通大學

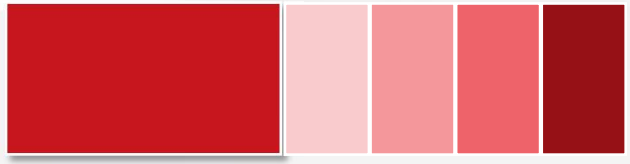
SHANGHAI JIAO TONG UNIVERSITY

感谢聆听

饮水思源 爱国荣校

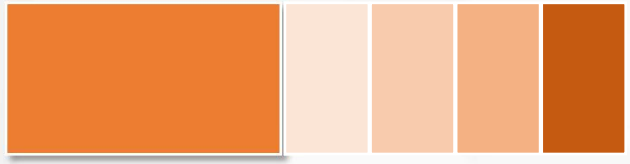
色彩规范

主色



对比色

平衡色



浅色与深色

建议尽量选择以上调色板中的颜色

字体规范

中文标题

微软雅黑

中文正文

微软雅黑

英文标题

Arial

英文正文

Arial

注意：

微软雅黑属于版权字体，商用请购买！

更多免费商用字体
https://jbox.sjtu.edu.cn/l/WuCIHQ

字体规范

中文标题	中文正文	中文标题	中文正文
微软雅黑	微软雅黑	思源黑体 Heavy 思源宋体 Heavy	思源黑体 Regular

英文标题	英文正文	英文标题	英文正文
Arial	Arial	Segoe UI	Segoe UI

注意：

微软雅黑属于版权字体，商用请购买！

更多免费商用字体
https://jbox.sjtu.edu.cn/l/WuCIHQ

了解更多思源系列字体
https://www.jianshu.com/p/d367f0f8a0f9

图标

图标