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

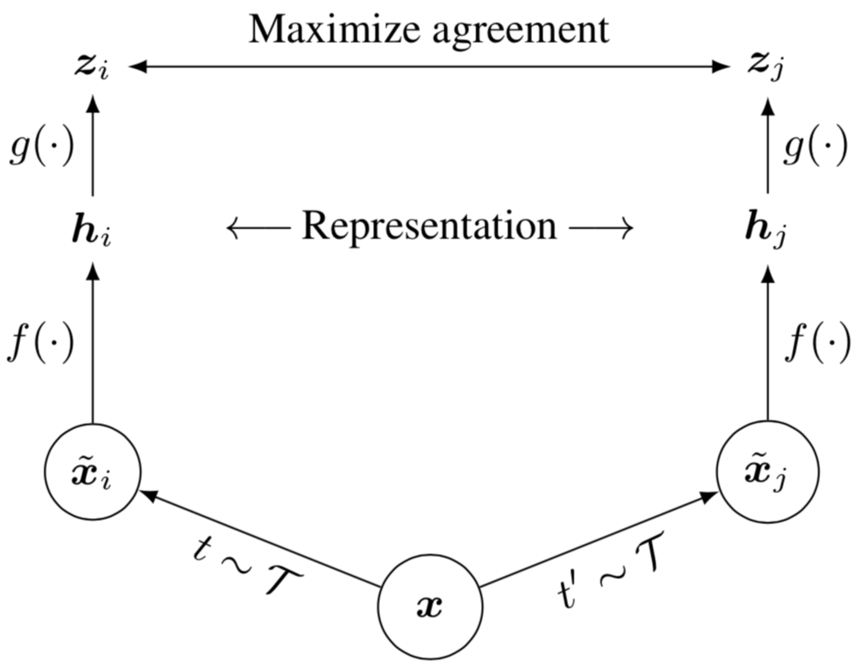
ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

Decoupled Contrastive Learning

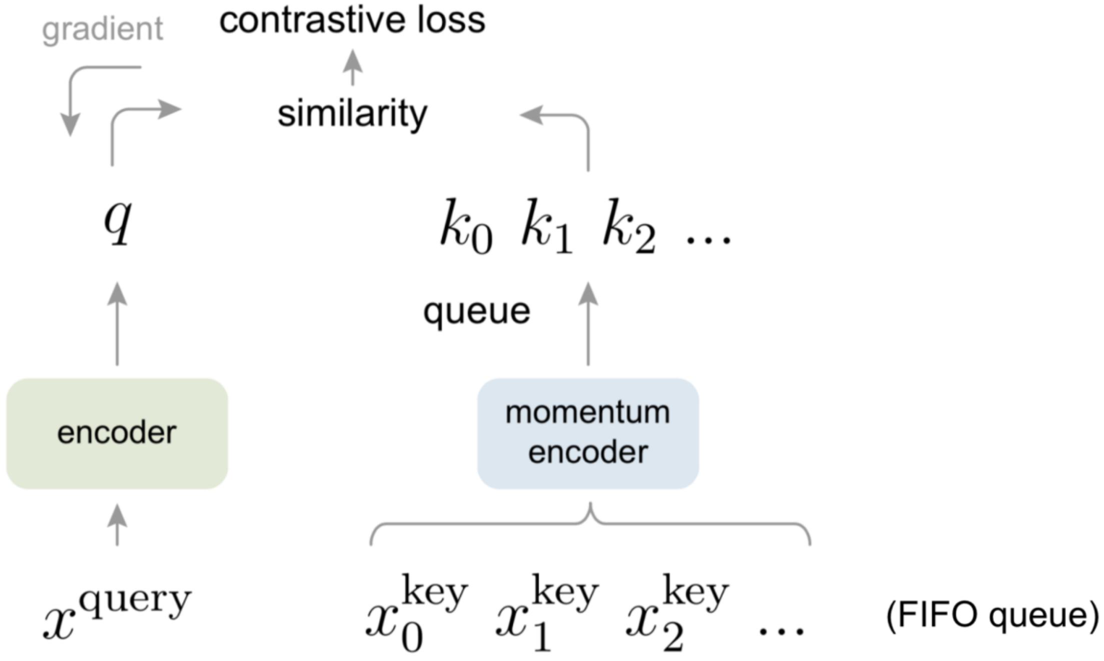
Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, Yann LeCun
IIS, Academia Sinica, Taiwan; UC Berkeley; National Taiwan University; Meta AI Research; New York University

ECCV 2022

Contrastive Learning

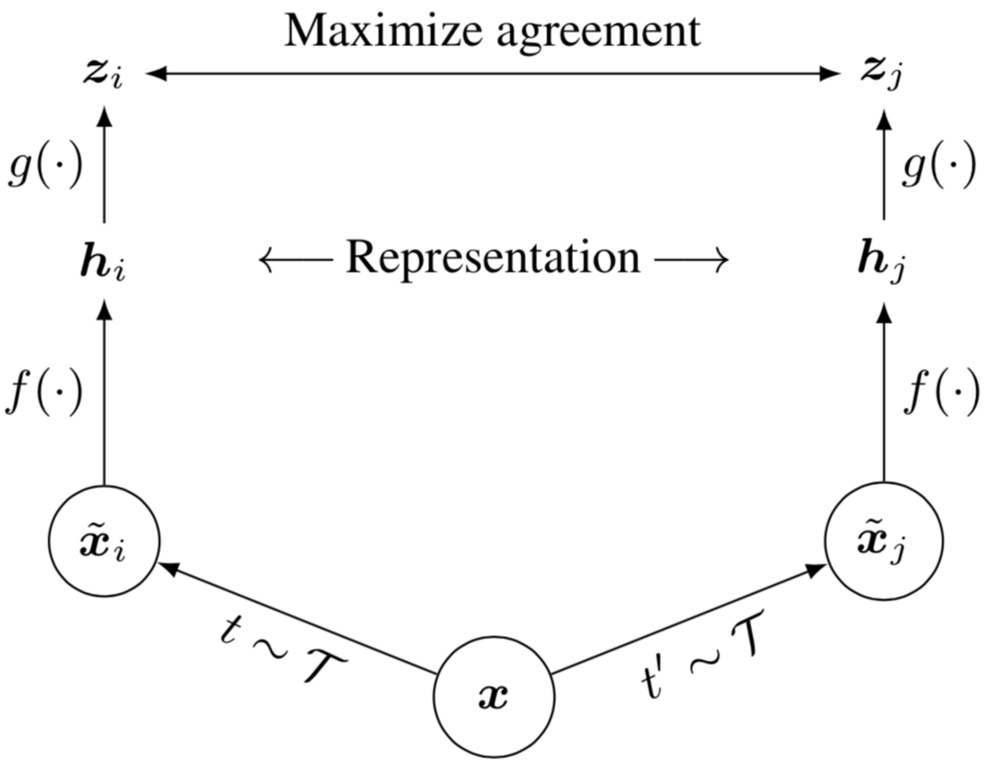


SimCLR



MoCo

SimCLR



$$L = \sum_{k\in\{1,2\},i\in[1,N]} L_i^{(k)}$$

where

$$L_i^{(k)} = -\log \frac{\exp\left(\left\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \right\rangle / \tau\right)}{\exp\left(\left\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \right\rangle / \tau\right) + U_{i,k}}$$

$$U_{i,k} = \sum_{l\in\{1,2\},j\in[1,N],j\neq i} \exp\left(\left\langle \mathbf{z}_i^{(k)}, \mathbf{z}_j^{(l)} \right\rangle / \tau\right)$$

Gradient

Take $L_i^{(1)}$ as an example:

$$L_i^{(1)} = -\log \frac{\exp\left(\left\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \right\rangle / \tau\right)}{\exp\left(\left\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \right\rangle / \tau\right) + U_{i,1}}$$

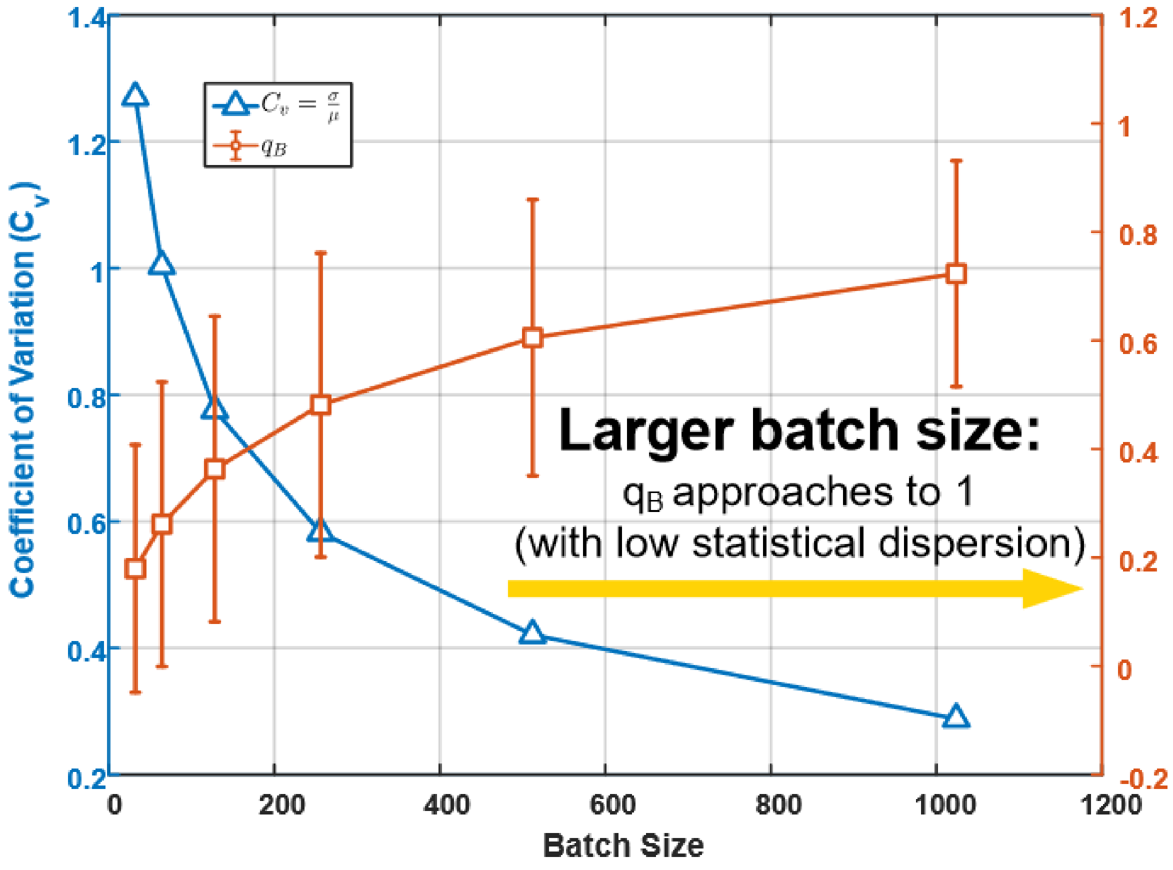
$$U_{i,1} = \sum_{l \in \{1,2\}, j \in [1,N], j \neq i} \exp\left(\left\langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(l)} \right\rangle / \tau\right)$$

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

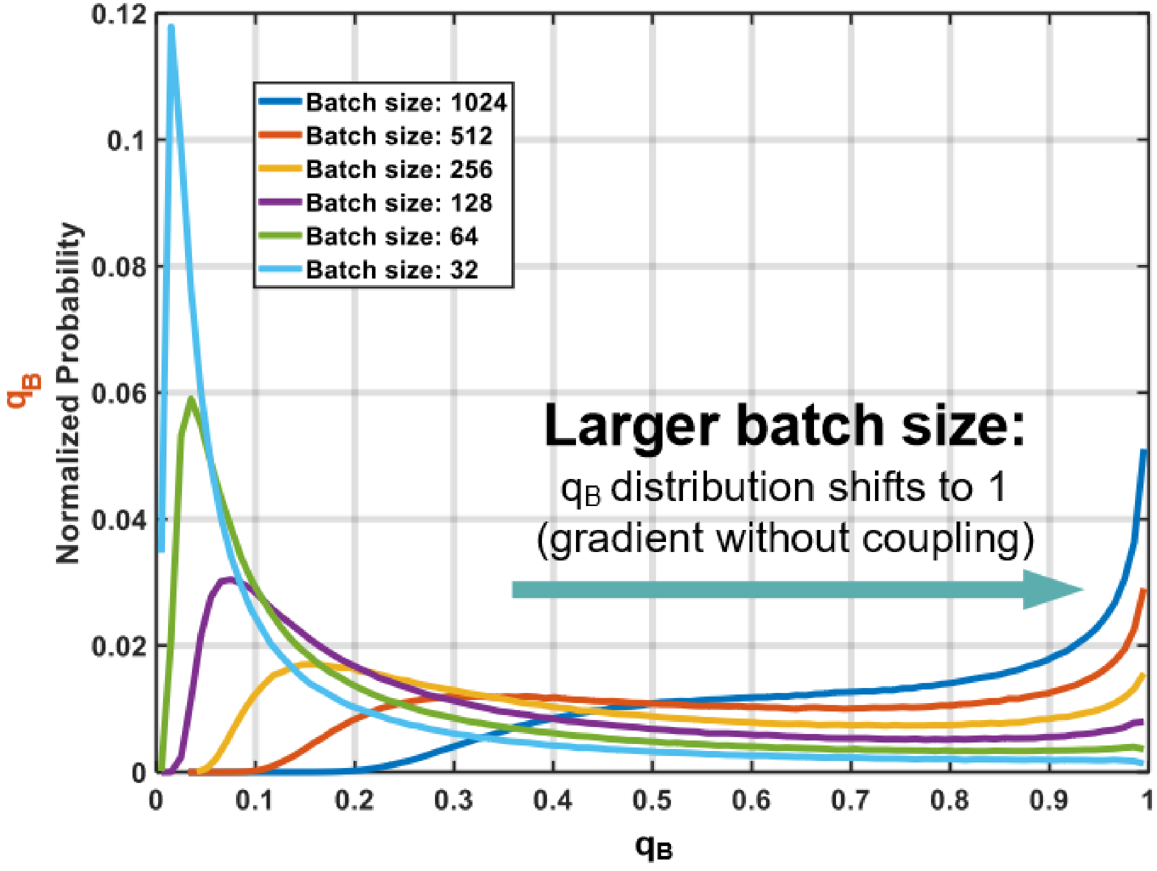
where $q_{B,i}^{(1)} = 1 - \frac{\exp\left(\left\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \right\rangle / \tau\right)}{\exp\left(\left\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \right\rangle / \tau\right) + U_{i,1}}$

Gradient

$$q_{B,i}^{(1)} = 1 - \frac{\exp\left(\left\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \right\rangle / \tau\right)}{\exp\left(\left\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \right\rangle / \tau\right) + U_{i,1}}$$



(a)



(b)

Decoupled Contrastive Learning

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

$$L_{DC,i}^{(k)} = -\log \frac{\exp\left(\left\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \right\rangle / \tau\right)}{U_{i,k}} = -\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \rangle / \tau + \log(U_{i,k})$$

Take $L_{DC,i}^{(1)}$ as an example:

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

Decoupled Contrastive Learning with Weight

$$L_{DCW,i}^{(k)} = -\omega(\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)})\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \rangle / \tau + \log(U_{i,k})$$

We want:

1. $E(\omega) = 1$

2. $E[\omega((\mathbf{z}_i^{(1)}, \mathbf{z}_j^{(l)}) \left\langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(l)} \right\rangle] \approx E[\left\langle \mathbf{z}_i^{(1)}, \mathbf{z}_j^{(l)} \right\rangle]$

3. less weight for more similar ones

The authors choose this function:

$$\omega(\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)}) = 2 - \frac{exp(\langle \mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)} \rangle / \sigma)}{\sum_j exp(\langle \mathbf{z}_j^{(1)}, \mathbf{z}_j^{(2)} \rangle / \sigma)}$$

Experiments

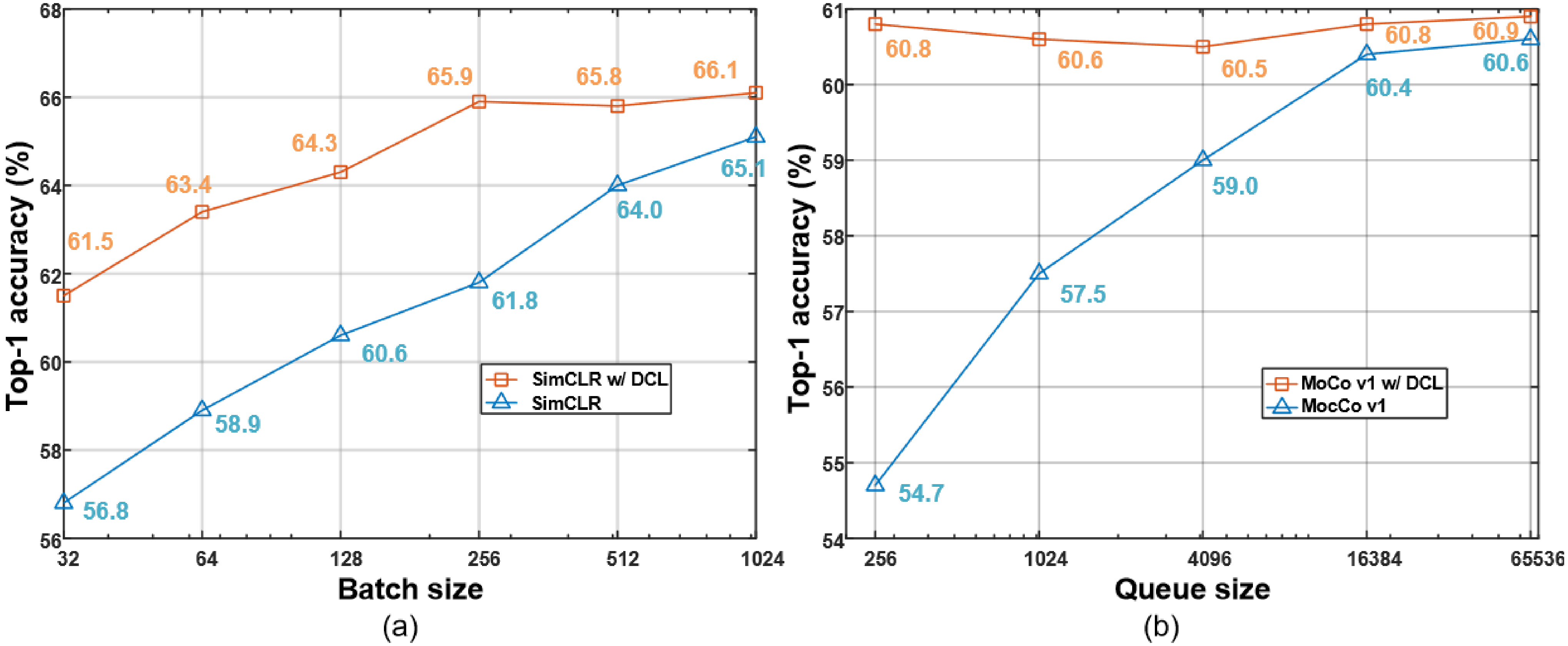


Fig. 3. Comparisons on ImageNet-1K with/without DCL under different numbers of (a): batch sizes for SimCLR and (b): queues for MoCo. Without DCL, the top-1 accuracy significantly drops when batch size (SimCLR) or queues (MoCo) becomes very small. Note that the temperature τ is 0.1 for SimCLR and 0.07 for MoCo in the comparison.

Experiments

Batch Size	32	64	128	256	512
Dataset	ImageNet-1K (kNN / Linear)				
Baseline (ResNet-50)	40.2/56.8	42.9/58.9	45.1/60.6	46.3/61.8	49.4/64.0
w/ DCL (ResNet-50)	**43.7/61.5**	**46.3/63.4**	**48.5/64.3**	**49.8/65.9**	**50.1/65.8**
Dataset	ImageNet-100 (kNN / Linear)				
Baseline (ResNet-50)	67.8/74.2	71.9/77.6	73.2/79.3	74.6/80.7	75.4/81.3
w/ DCL (ResNet-50)	**74.9/80.8**	**76.3/82.0**	**76.5/81.9**	**76.9/83.1**	**76.8/82.8**
Dataset	CIFAR10 (kNN / Linear)				
Baseline (ResNet-18)	78.9/79.8	80.4/81.3	81.1/82.8	81.4/83.0	81.3/83.3
w/ DCL (ResNet-18)	**83.7/85.1**	**84.4/85.9**	**84.4/85.7**	**84.2/85.3**	**83.5/84.7**
Dataset	CIFAR100 (kNN / Linear)				
Baseline (ResNet-18)	49.4/51.3	50.3/53.8	51.8/55.3	52.0/56.3	52.4/56.8
w/ DCL (ResNet-18)	**51.1/55.4**	**54.3/58.3**	**54.6/58.9**	**54.9/58.5**	**55.0/58.4**
Dataset	STL10 (kNN / Linear)				
Baseline (ResNet-18)	74.1/76.2	77.6/77.8	79.3/80.0	80.7/81.3	81.3/81.5
w/ DCL (ResNet-18)	**82.0/85.2**	**82.8/86.3**	**81.8/86.1**	**81.2/85.7**	**81.0/85.6**

Experiments

Table 2. Comparisons between SimCLR baseline, DCL, and DCLW. The linear and kNN top-1 (%) results indicate that DCL improves baseline performance, and DCLW further provides an extra boost. Note that results are under batch size 256 and epoch 200. All models are both trained and evaluated with the same experimental settings. The backbones are ResNet-18 and ResNet-50 for CIFAR and ImageNet, respectively.

Dataset	CIFAR10 (kNN)	CIFAR100 (kNN)	ImageNet-100 (linear)	ImageNet-1K (linear)
SimCLR	81.4	52.0	80.7	61.8
DCL	84.2 (+**2.8**)	54.9 (+**2.9**)	83.1 (+**2.4**)	65.9 (+**4.1**)
DCLW	**84.8** (+**3.4**)	**55.2** (+**3.2**)	**84.2** (+**3.5**)	**66.9** (+**5.1**)

Experiments

Table 4. The comparisons with/without DCL under various batch sizes from 32 to 512 on ResNet-50.

Architecture@epoch	ResNet-50@500 epoch									
Dataset	CIFAR10 (kNN)					CIFAR100 (kNN)				
Batch Size	32	64	128	256	512	32	64	128	256	512
SimCLR	82.2	85.9	88.5	88.9	89.1	49.8	55.3	59.9	60.6	61.1
SimCLR w/ DCL	**86.1**	**88.3**	**89.9**	**90.1**	**90.3**	**54.3**	**58.4**	**61.6**	**62.0**	**62.2**

Table 5. Linear top-1 accuracy (%) comparison with MoCo-V2 on ImageNet-1K and ImageNet-100.

Queue Size	32	64	128	256	8192	64	256	65536
Dataset	ImageNet-100 (Linear)					ImageNet-1K (Linear)		
MoCo-v2 Baseline (ResNet-50)	73.7	76.4	78.7	78.7	79.8	63.9	67.1	67.5
MoCo-v2 w/DCL (ResNet-50)	**76.2**	**78.3**	**79.6**	**79.6**	**80.5**	**65.8**	**67.6**	**67.7**

Experiments

Table 6. ImageNet-1K top-1 accuracy (%) on SimCLR and MoCo-v2 with/without DCL under few training epochs. We further list results under 200 epochs for clear comparison. With DCL, the performance of SimCLR trained under 100 epochs nearly reaches its performance under 200 epochs. The MoCo-v2 with DCL also reaches higher accuracy than the baseline under 100 epochs.

	SimCLR	SimCLR w/ DCL	MoCo-v2	MoCo-v2 w/ DCL
100 Epoch	57.5	64.6	63.6	64.4
200 Epoch	61.8	65.9	67.5	67.7

Experiments

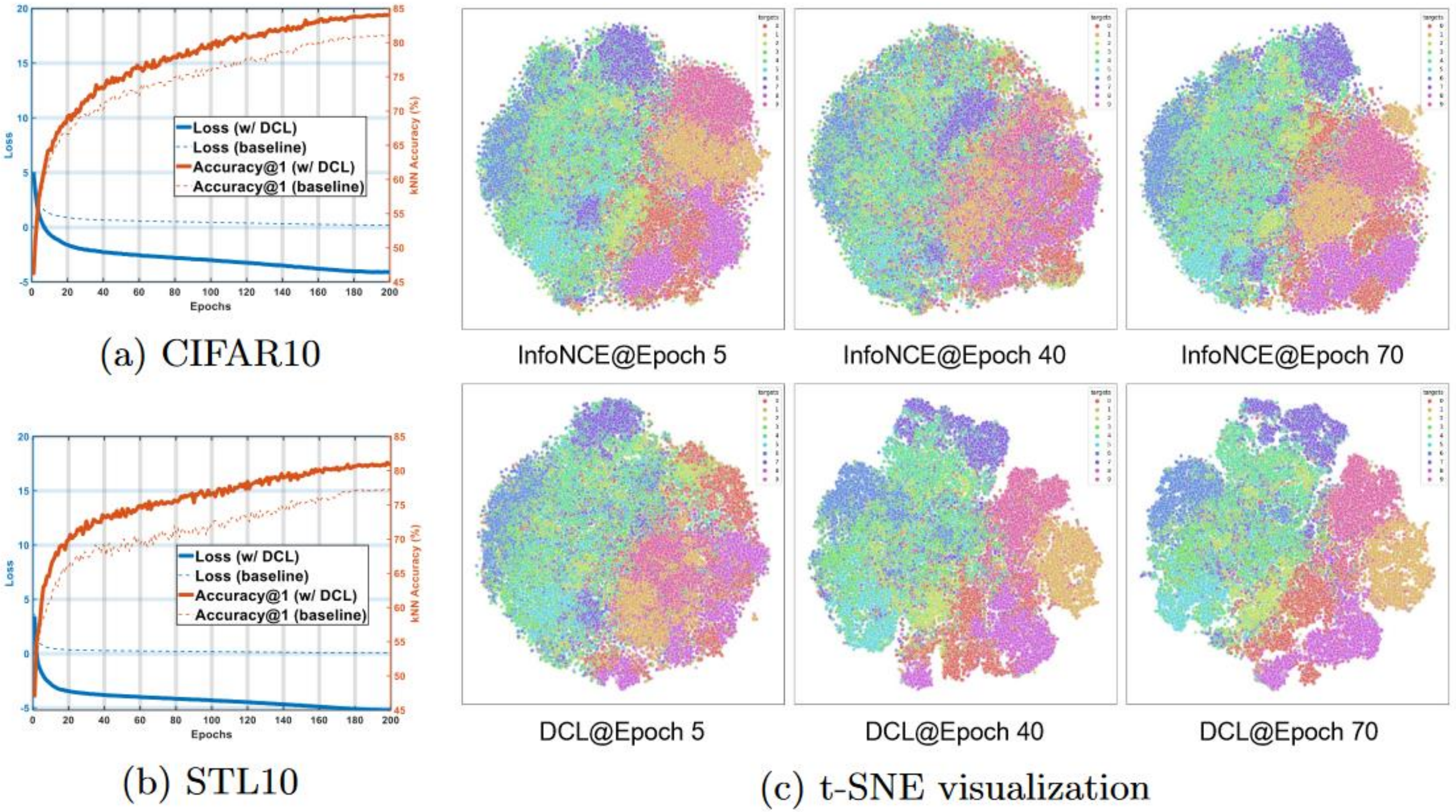


Fig. 4. Comparisons between DCL and InfoNCE-based baseline (SimCLR) on (a) CIFAR10 and (b) STL10 data. DCL speeds up the model convergence during the SSL pre-training and provides better performance than the baseline on CIFAR and STL10 data. (c) t-SNE visualization of CIFAR10 with 32 batch size. DCL shows a stronger separation force between the features than SimCLR.