Bootstrap Your Own Latent
A New Approach to Self-Supervised Learning

2023. 05. 11

BYOL (2020, Neural Information Processing System)

❖ Bootstrap Your Own Latent A New Approach to Self-Supervised Learning

- 2020년 Neural Information Processing System에 게재되었으며, 3295회 인용됨
- 대표적인 Non-contrastive learning 방법론으로 negative sample을 사용하지 않고 SOTA 달성

Bootstrap Your Own Latent
A New Approach to Self-Supervised Learning

Jean-Bastien Grill[*1] , **Florian Strub**[*1] , **Florent Altché**[*1] , **Corentin Tallec**[*1] , **Pierre H. Richemond**[*1,2]
Elena Buchatskaya[1] , **Carl Doersch**[1] , **Bernardo Avila Pires**[1] , **Zhaohan Daniel Guo**[1]
Mohammad Gheshlaghi Azar[1], **Bilal Piot**[1], **Koray Kavukcuoglu**[1] , **Rémi Munos**[1] , **Michal Valko**[1]

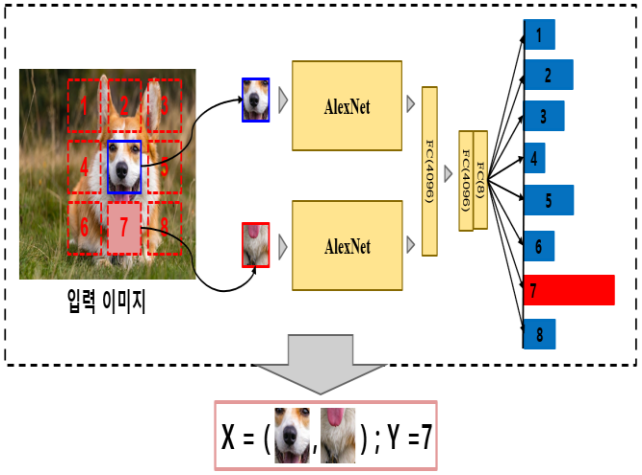
[1]DeepMind [2]Imperial College

[jbgrill,fstrub,altche,corentint,richemond]@google.com

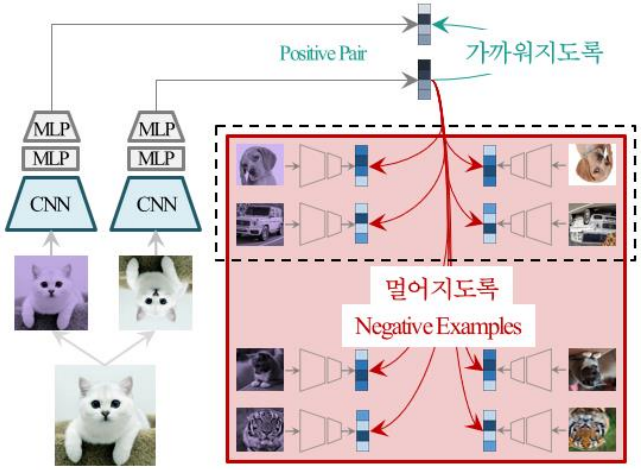
Background

❖ 기존 자기 지도학습 방법의 한계점

- Pretext task
  - ➢ 모델의 성능이 사전학습을 위해 사용되는 pretext task의 질에 따라 차이가 심하다는 한계가 존재
- Contrastive Learning: Pretext task의 한계를 개선한 방법
  1. 데이터 증강 기법에 따른 성능 편차가 심함
  2. 방대한 양의 배치 크기가 필요함 (많은 컴퓨팅 자원이 요구됨)
  3. Negative Pair 선정이 어려움 → Supervised Contrastive Learning (클래스 레이블을 활용해서 negative sample 내에 false negative를 positive sample로 변환



Context Prediction

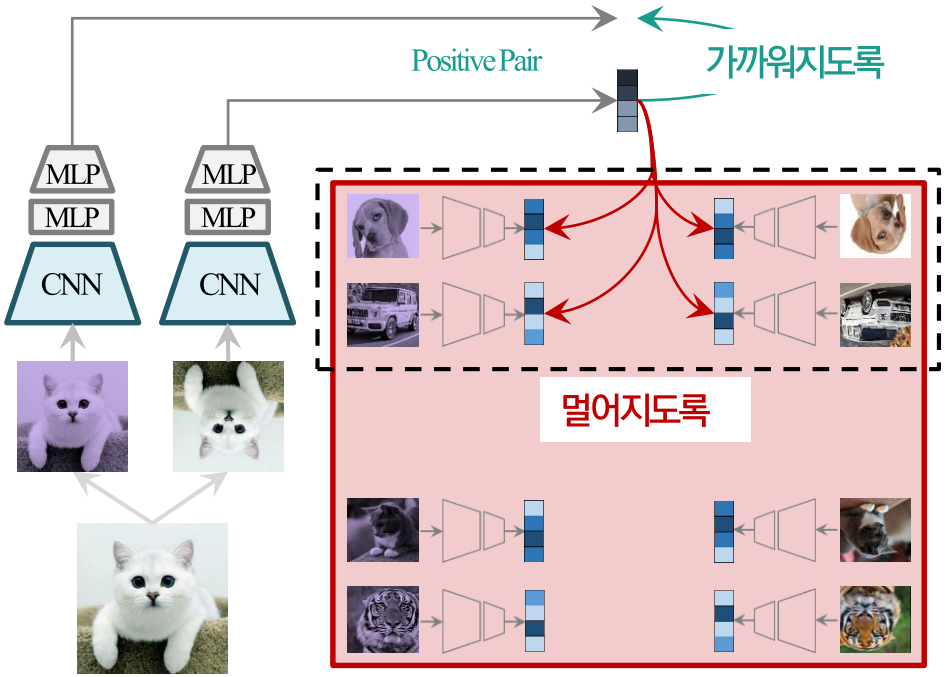


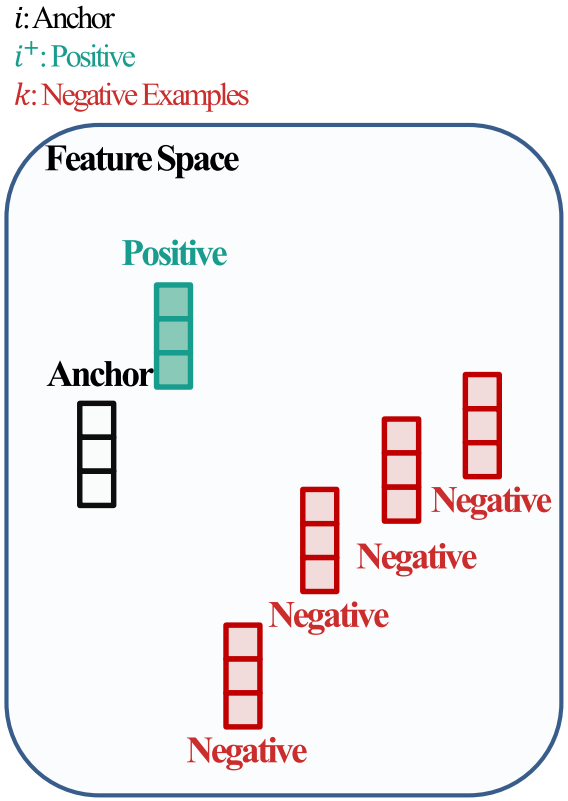
SimCLR

Background

❖ 대조 학습의 한계점

1. 데이터 증강 기법에 따른 성능 편차가 심함 → augmented negative sample = positive sample
2. 방대한 양의 배치 크기가 필요함 (많은 컴퓨팅 자원이 요구됨) → because of negative sample
3. Negative Pair 선정이 어려움 → anchor에 augmentation을 취한 것은 positive, 다른 샘플은 negative ➔ false negative





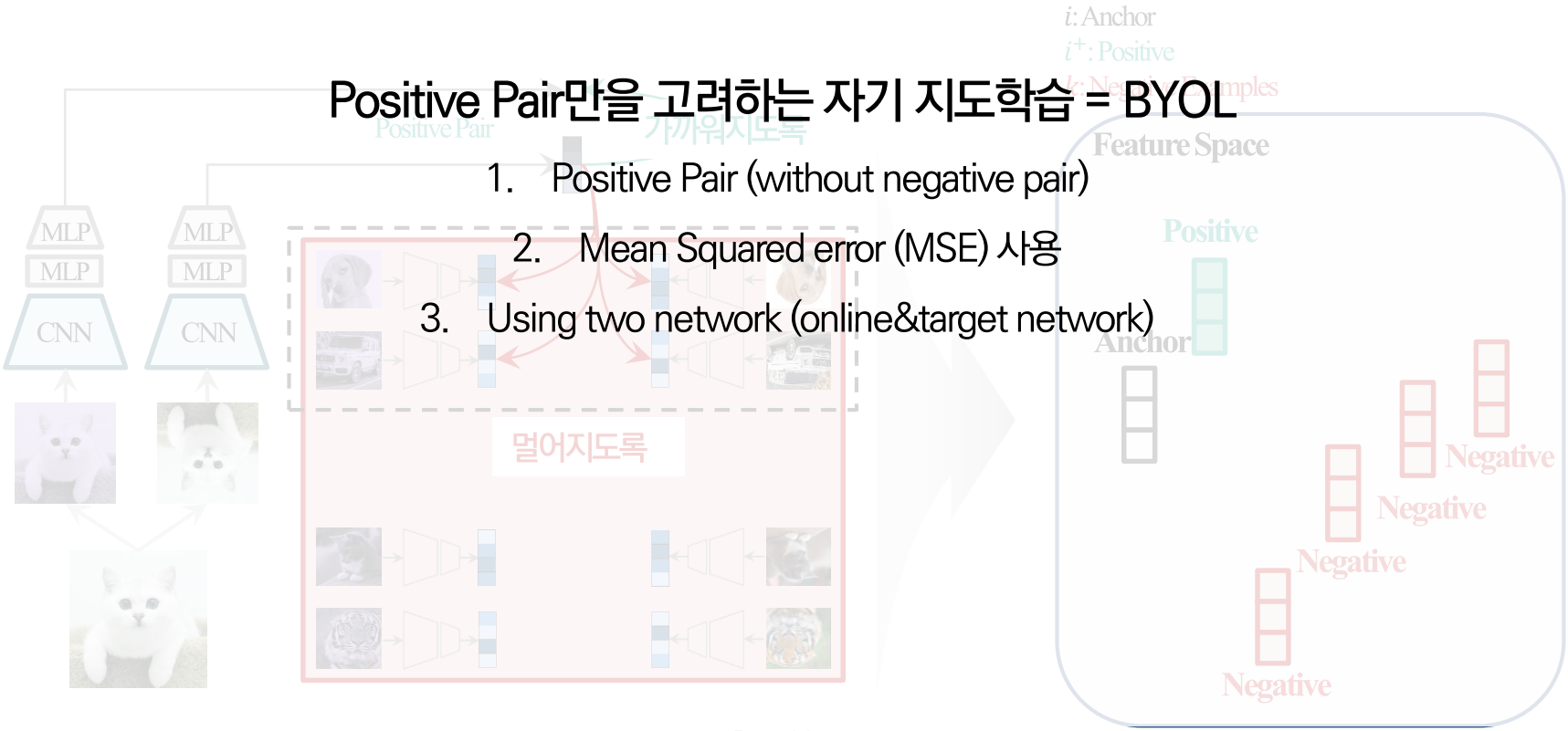
Background

❖ 대조 학습의 한계점

1. 데이터 증강 기법에 따른 성능 편차가 심함 → augmented negative sample = positive sample
2. 방대한 양의 배치 크기가 필요함 (많은 컴퓨팅 자원이 요구됨) → because of negative sample
3. Negative Pair 선정이 어려움 → anchor에 augmentation을 취한 것은 positive, 다른 샘플은 negative → false negative

Positive Pair만을 고려하는 자기 지도학습 = BYOL

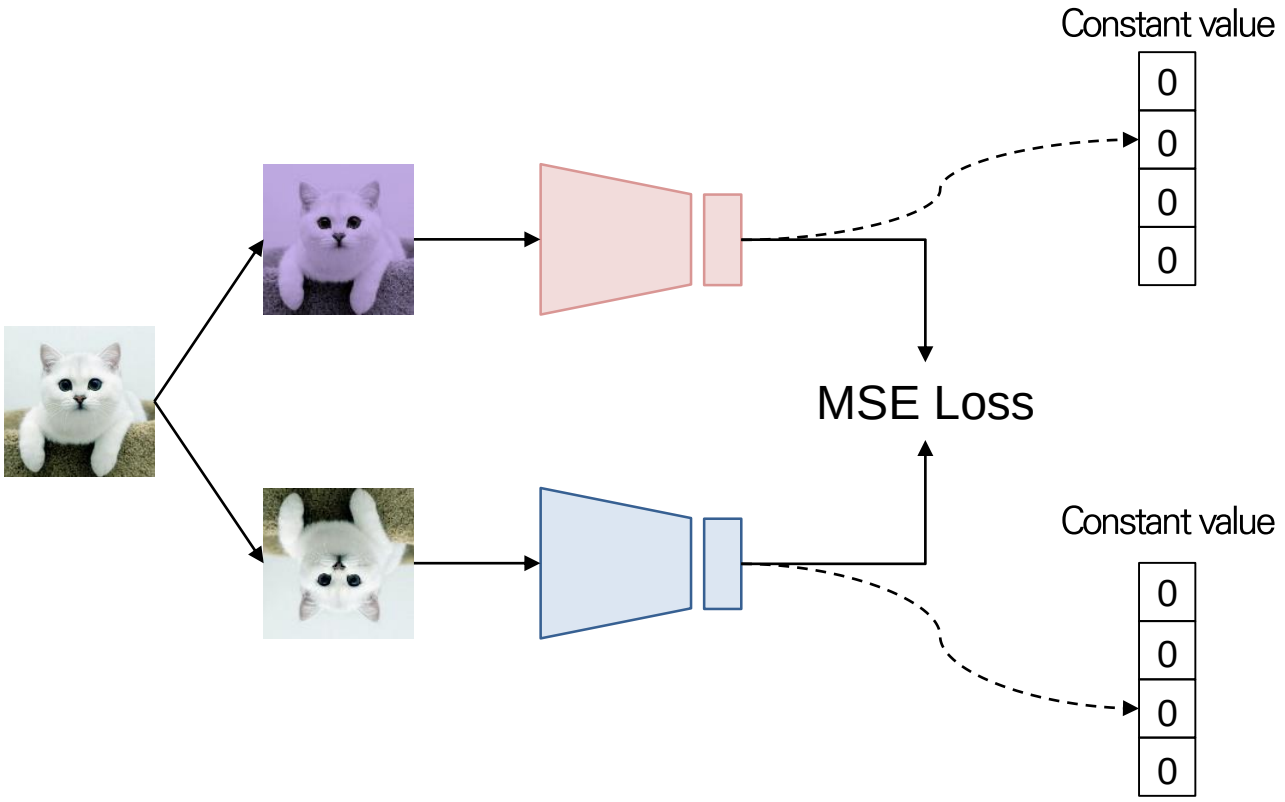
1. Positive Pair (without negative pair)
2. Mean Squared error (MSE) 사용
3. Using two network (online&target network)



Background

❖ Collapsed representation problem

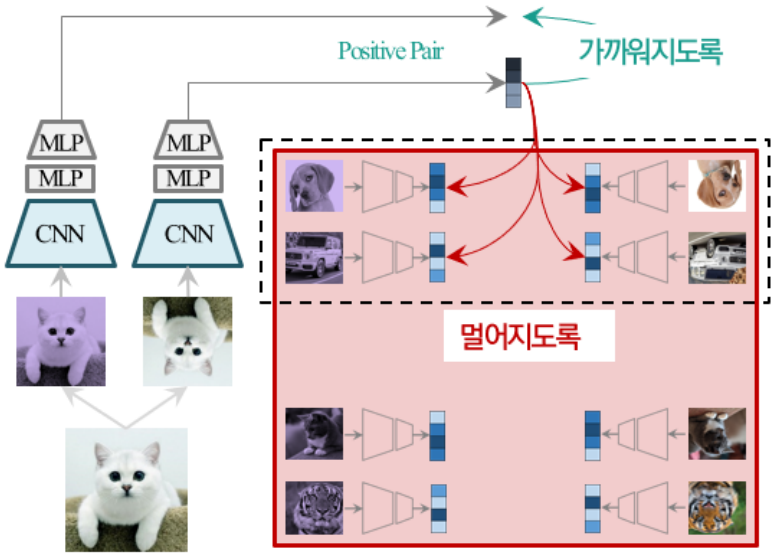
- 동일한 네트워크를 사용하였을 때, 학습이 전혀 안되고 동일한 값을 출력하는 문제



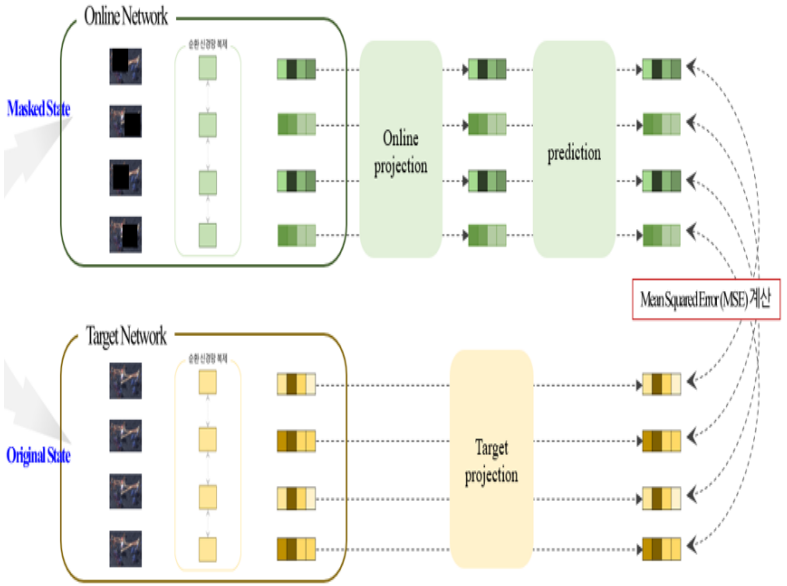
Background

❖ Collapsed representation problem

- Contrastive Learning: negative sample을 사용해 해당 문제 해결
- Non-contrastive Learning: 두 개의 network를 사용하여 해당 문제 해결



SimCLR

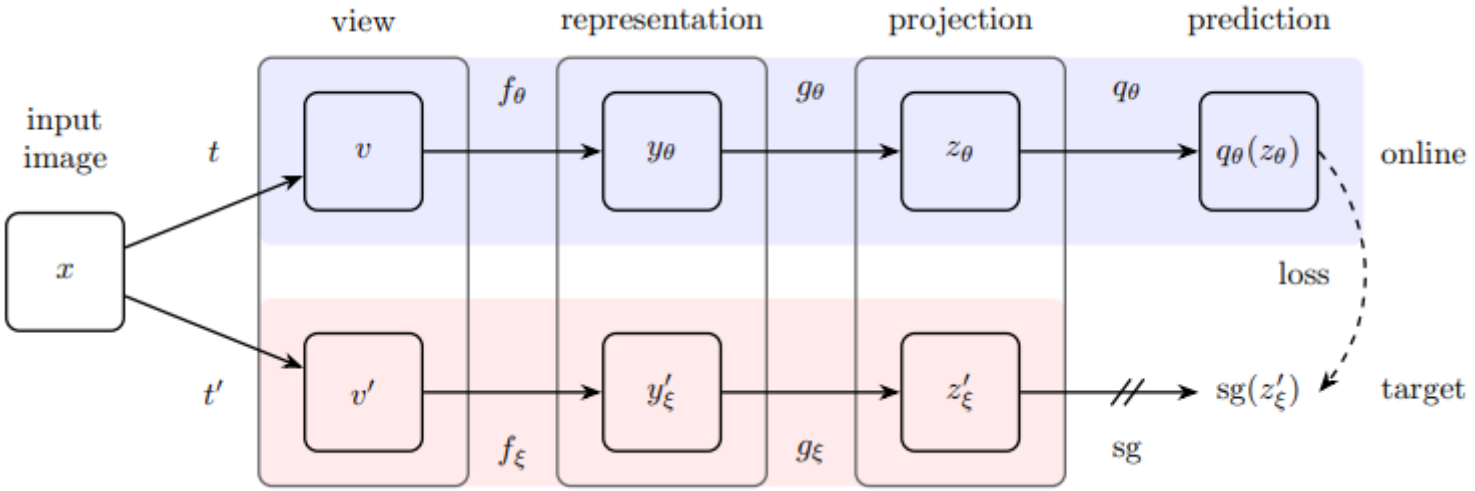


BYOL

Proposed Method

❖ Overall Architecture

- 서로 다른 파라미터를 갖는 online network, target network 존재
  - ➢ Target network의 output은 online network의 output의 정답으로 활용
- Online network update: MSE loss를 구해 gradient update 진행
- Target network update: exponential moving average로 진행 ($\xi \leftarrow \tau\xi + (1-\tau)\xi$)



$$\textbf{Total loss: } L_\theta^{BYOL} + \tilde{L}_\theta^{BYOL}$$

Proposed Method

❖ Overall Architecture

- Online network의 prediction과 target network의 projection에 L2 Regularization을 취한 뒤 L_θ^{BYOL} 계산

$$\mathcal{L}_{\theta,\xi} \triangleq \left\| \overline{q_\theta}(z_\theta) - \overline{z}'_\xi \right\|_2^2 = 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{\left\| q_\theta(z_\theta) \right\|_2 \cdot \left\| z'_\xi \right\|_2}.$$

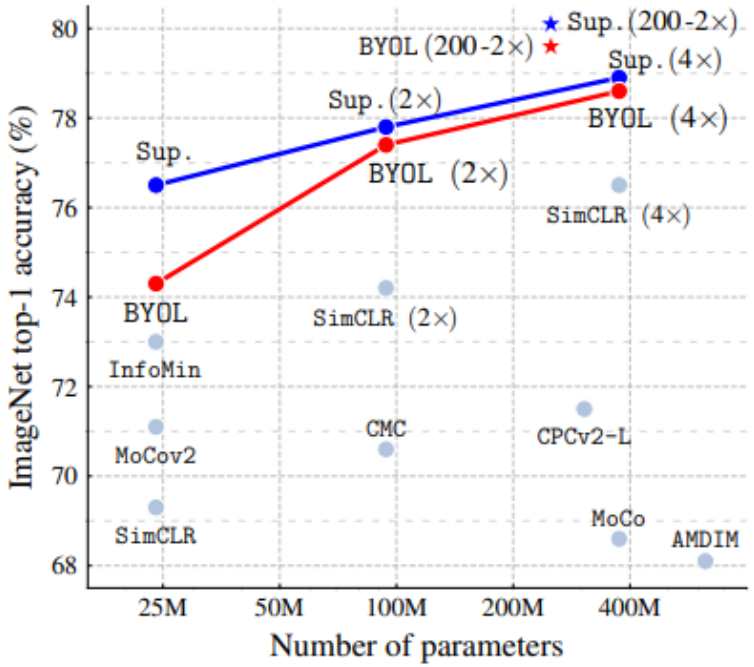
- Total Loss는 online network, target network에서 사용된 augmentation 방법을 교환하여 얻은 loss와 기존 loss의 합을 사용

$$Total\ loss\colon\ L_\theta^{BYOL} + \tilde{L}_\theta^{BYOL}$$

Experiment

❖ Experiment 1

- ImageNet 데이터 셋을 활용해 classification을 했을 때 성능 (Encoder는 freeze하고 진행)
  - ✓ 표준 ResNet-50을 사용했을 때, 74.3% 달성 (top-1 accuracy)
  - ✓ 더 거대한 ResNet을 사용했을 때, 79.6% 달성 (top-1 accuracy)



Experiment

❖ Experiment 2

Method	Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
BYOL (ours)	**75.3**	91.3	**78.4**	**57.2**	**62.2**	**67.8**	60.6	82.5	75.5	90.4	94.2	**96.1**
SimCLR (repro)	72.8	90.5	74.4	42.4	60.6	49.3	49.8	81.4	**75.7**	84.6	89.3	92.6
SimCLR [8]	68.4	90.6	71.6	37.4	58.8	50.3	50.3	80.5	74.5	83.6	90.3	91.2
Supervised-IN [8]	72.3	**93.6**	78.3	53.7	61.9	66.7	**61.0**	**82.8**	74.9	**91.5**	**94.5**	94.7
Fine-tuned:												
BYOL (ours)	**88.5**	**97.8**	86.1	**76.3**	63.7	91.6	**88.1**	**85.4**	**76.2**	91.7	**93.8**	97.0
SimCLR (repro)	87.5	97.4	85.3	75.0	63.9	91.4	87.6	84.5	75.4	89.4	91.7	96.6
SimCLR [8]	88.2	97.7	85.9	75.9	63.5	91.3	88.1	84.1	73.2	89.2	92.1	97.0
Supervised-IN [8]	88.3	97.5	**86.4**	75.8	**64.3**	**92.1**	86.0	85.0	74.6	**92.1**	93.3	**97.6**
Random init [8]	86.9	95.9	80.2	76.1	53.6	91.4	85.9	67.3	64.8	81.5	72.6	92.0

Table 3: Transfer learning results from ImageNet (IN) with the standard ResNet-50 architecture.

Method	AP_{50}	mIoU
Supervised-IN [9]	74.4	74.4
MoCo [9]	74.9	72.5
SimCLR (repro)	75.2	75.2
BYOL (ours)	**77.5**	**76.3**

(a) Transfer results in semantic segmentation and object detection.

Method	Higher better			Lower better	
	pct.< 1.25	pct.$< 1.25^2$	pct.$< 1.25^3$	rms	rel
Supervised-IN [83]	81.1	95.3	98.8	0.573	**0.127**
SimCLR (repro)	83.3	96.5	99.1	0.557	0.134
BYOL (ours)	**84.6**	**96.7**	**99.1**	**0.541**	0.129

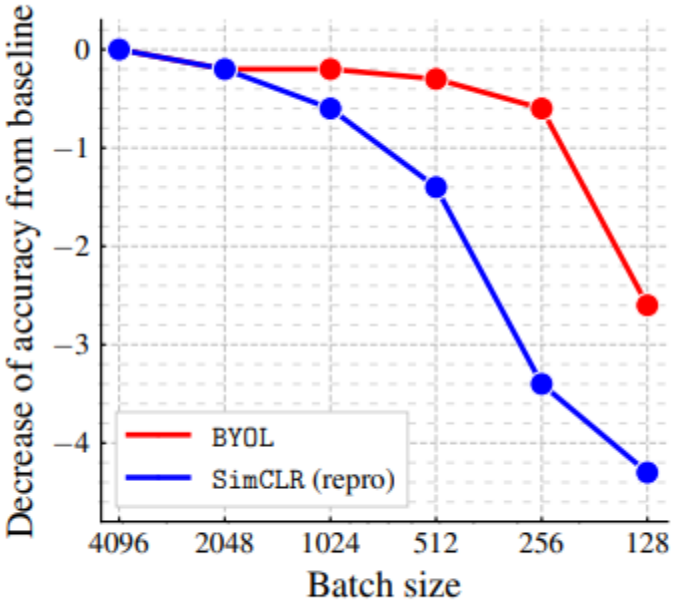
(b) Transfer results on NYU v2 depth estimation.

Experiment

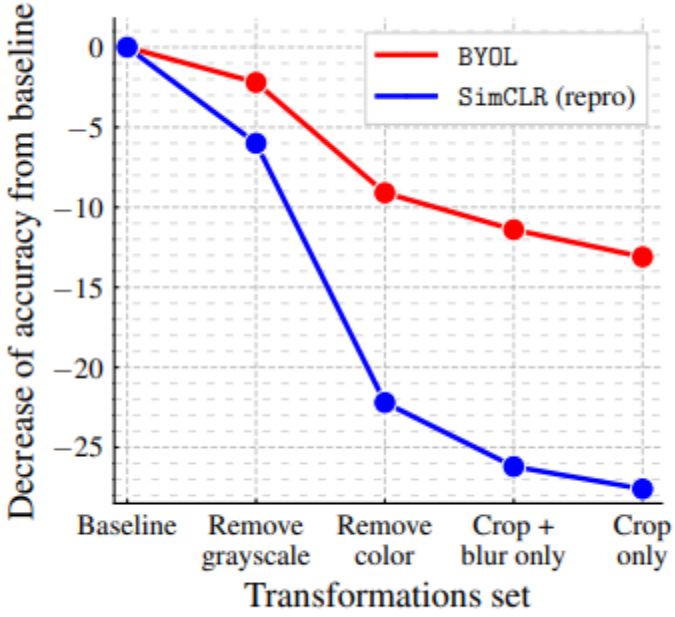
❖ Experiment 3

a. Batch size에 따른 contrastive learning과 non-contrastive learning 성능 비교 실험 결과

b. Augmentation에 사용 유무에 따른 강건성 비교 실험 결과



(a) Impact of batch size



(b) Impact of progressively removing transformations