

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

DeepMind, Imperial College

NIPS 2020

0. Index

- ◆ Introduction
- ◆ Design
- ◆ Evaluation
- ◆ Conclusion

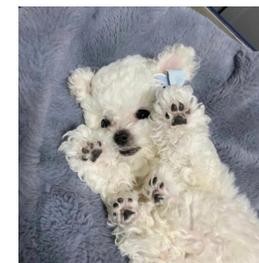
01. Introduction

01. Introduction : Representation Learning

How to Solve Cats VS Dogs problem?



01. Introduction : Representation Learning



01. Introduction : Representation Learning



01. Introduction : Representation Learning

 $=$  \neq 

01. Introduction : Contrastive Learning

◆ Contrastive Learning

- 유사한 이미지가 저차원 공간에서 서로 가깝게, 다른 이미지는 서로 멀리 떨어져 있도록 저차원 공간에서 이미지를 인코딩하는 방법을 모델이 학습하는 것을 의미

01. Introduction : Contrastive Learning

◆ Contrastive Learning의 단점

- Require careful treatment of negative pairs
 - ✓ Negative pairs를 제공하는 전략을 고려할 필요 있음
ex) SimCLR : Batch Size
- Choice of Image Augmentation
 - ✓ Augmentation 조합에 따라 모델의 성능이 크게 좌우됨

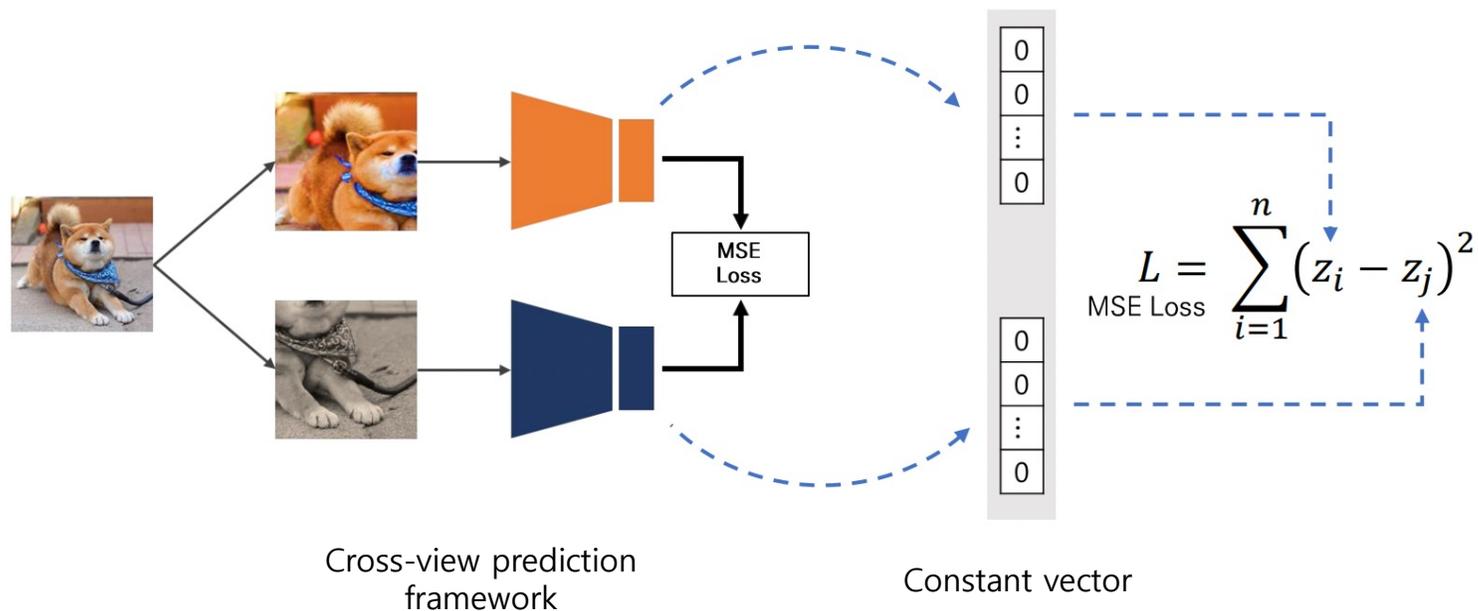


- Requires comparing each representation of an augmented view with many negative example

01. Introduction : Contrastive Learning

◆ Collapsed representation (Mode collapse)

- Positive pairs로만 학습을 하는 경우 모델이 constant vector만을 출력하는 문제
 - ✓ Train loss는 작아지지만 학습은 전혀 안되는 문제 발생



01. Introduction : Contrastive Learning

◆ Collapsed representation (Mode collapse)

- Contrastive loss는 positive와 negative sample을 모두 사용하여 collapse를 방지
 - ✓ Positive pair간의 유사도가 크고 Negative pair간의 유사도가 작을수록 loss값이 작아짐

$$L_{i,j} = -\log \frac{\exp\left(\frac{\text{sim}(z_i, z_j)}{\tau}\right)}{\sum_{k=1}^N [k \neq i] \exp\left(\frac{\text{sim}(z_i, z_k)}{\tau}\right)}$$

Contrastive Loss

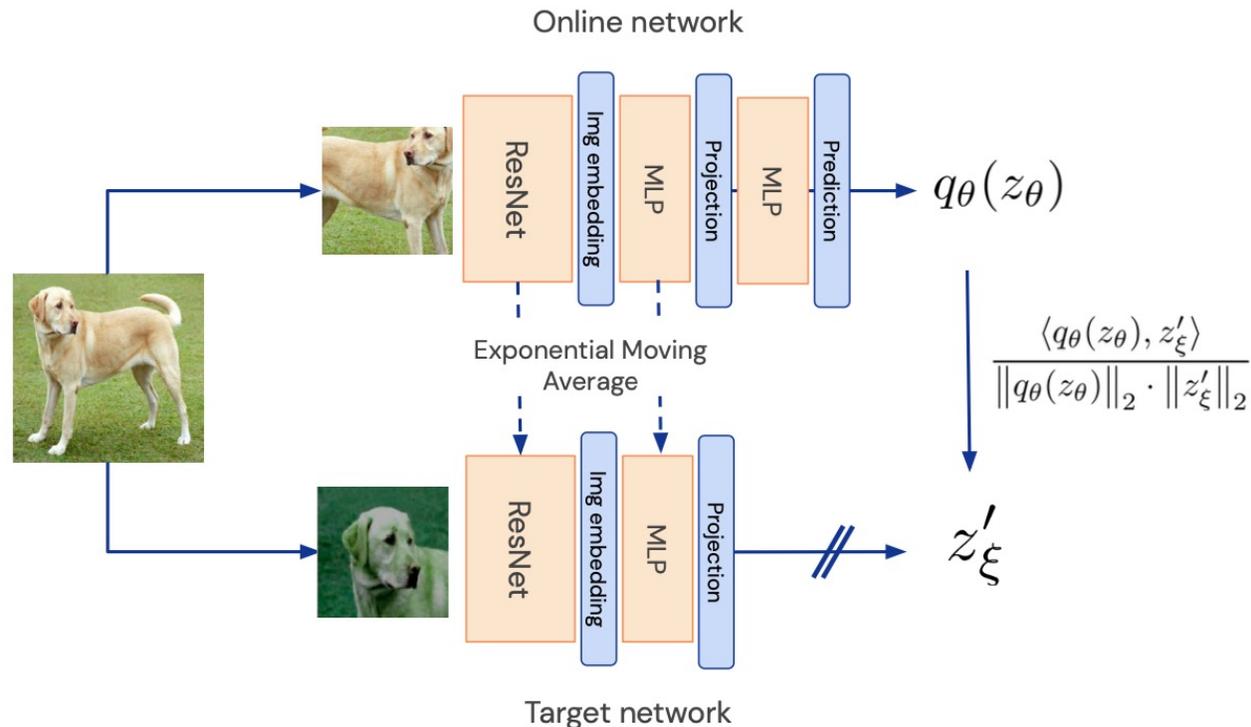
Cosine similarity (Positive pair)

Cosine similarity (Negative pair)

01. Introduction : BYOL

◆ Bootstrap Your Own Latent (BYOL)

- 기존 Contrastive learning에서 negative sample에 의존하는 방식을 벗어나고자 함
- Positive sample만으로 mode collapse 현상 없이 representation learning 가능



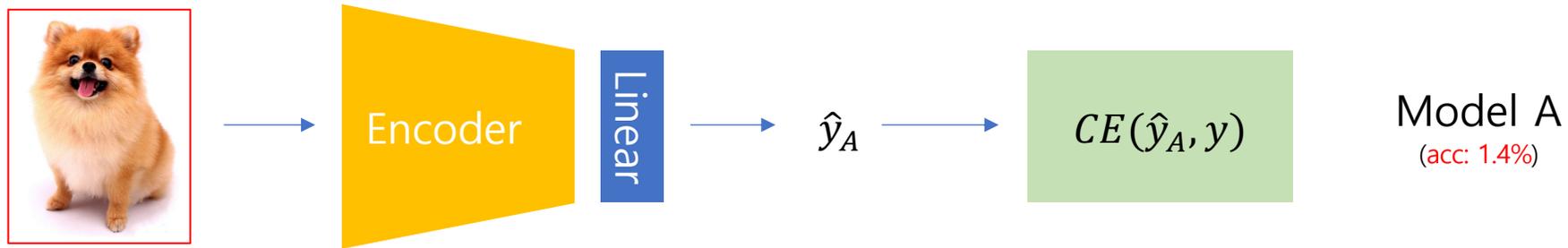
02. Design

02. Design: Overview

◆ Core Motivation of BYOL

- BYOL의 Motive가 된 간단한 실험

✓ Case 1) Encoder(random parameter / freeze) + MLP

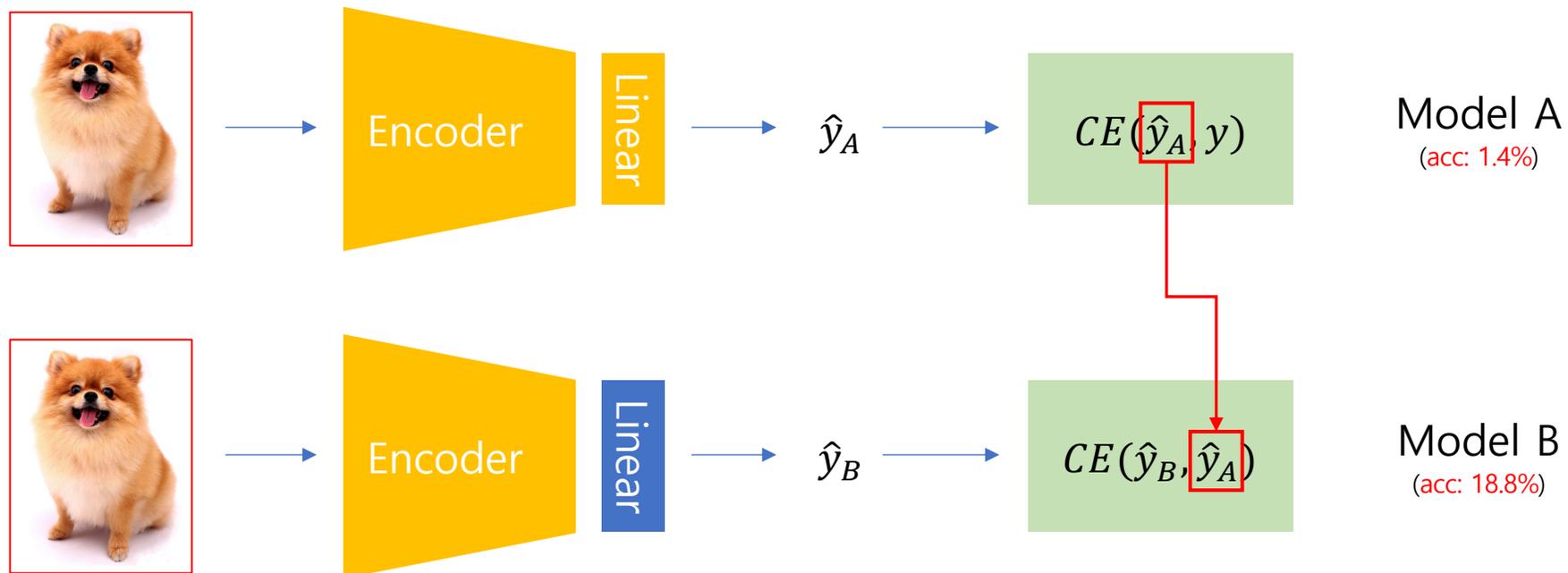


02. Design: Overview

◆ Core Motivation of BYOL

- BYOL의 Motive가 된 간단한 실험

✓ Case 2) Case 1과 동일한 구조의 네트워크를 생성하여 Case 1의 네트워크가 출력한 값을 예측하도록 학습

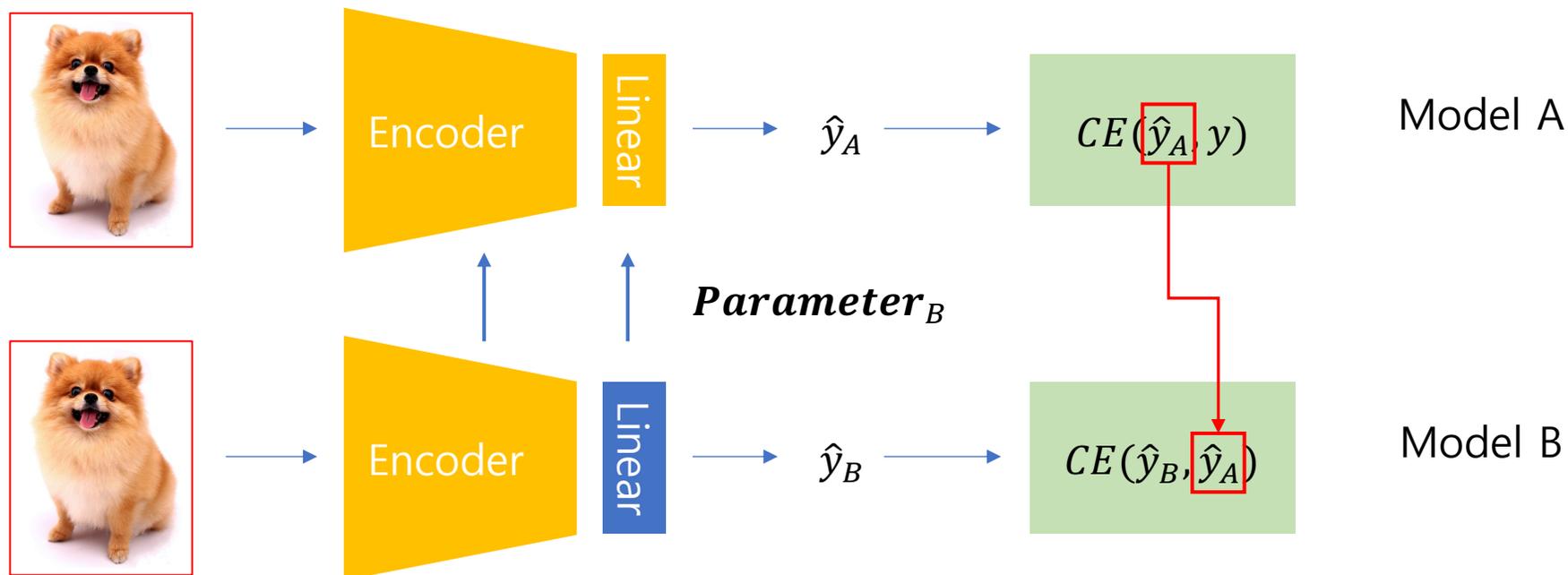


02. Design: Overview

◆ Core Motivation of BYOL

- BYOL의 Motive가 된 간단한 실험

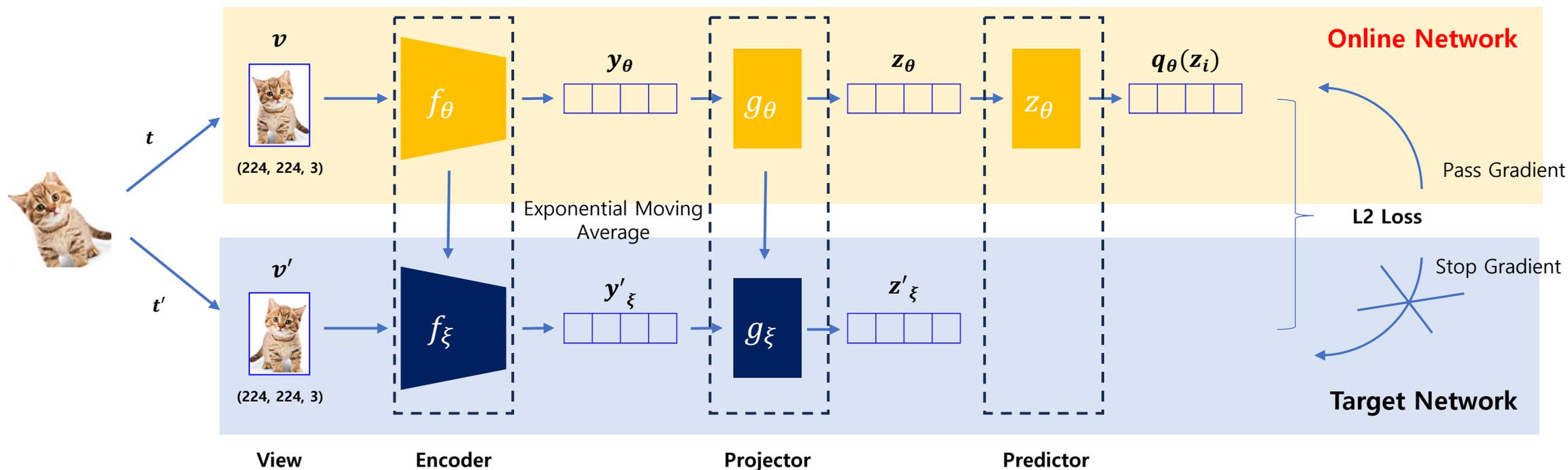
✓ if) 만약 Model B의 파라미터를 이용하여 타겟이 되는 Model A의 파라미터를 update한다면? (Bootstrap)



02. Design: Model Structure

◆ Architecture of BYOL

- 파라미터를 업데이트하는 방식이 서로 다른 동일한 구조의 두 네트워크로 구성됨 (Online Network / Target Network)
- 두 네트워크 모두 Encoder와 Projector를 보유하나 Predictor는 Online Network에서만 보유함
- Target Network에서 출력한 representation vector를 Online Network에서 예측하는 훈련을 진행함



02. Design: Model Structure

◆ Target Network update in BYOL

- Exponential moving average
 - ✓ Online Network의 weight를 이용하여 Target Network의 weight를 점진적으로 update하는 방식
 - ✓ Cosine annealing을 사용하여 학습이 진행될수록 τ 를 점점 1에 가까운 값으로 키움

$$\underbrace{\xi}_{\text{Target Network (New)}} \leftarrow \underbrace{\tau \xi}_{\text{Target Network (Old)}} + \underbrace{(1 - \tau)}_{\substack{\text{Online Network} \\ \uparrow \\ \tau_{base} = 0.996}} \theta$$

$$\tau \triangleq 1 - (1 - \tau_{base}) \cdot \frac{\left(\cos \frac{\pi k}{K} + 1\right)}{2}$$

K : Maximum Number of Training step

k : Current Training step

02. Design: Loss Function

◆ Loss Function of BYOL

- L2 loss

- ✓ 각 네트워크의 Prediction과 Projection에 L2 정규화를 취한 뒤 loss를 계산

$$\mathcal{L}_{\theta,\xi} \triangleq \|\overline{q_{\theta}(z_{\theta})} - \overline{z'_{\xi}}\|_2^2$$

$$= 2 - 2 \cdot \frac{\langle q_{\theta}(z_{\theta}), z'_{\xi} \rangle}{\|q_{\theta}(z_{\theta})\|_2 \cdot \|z'_{\xi}\|_2}$$

←

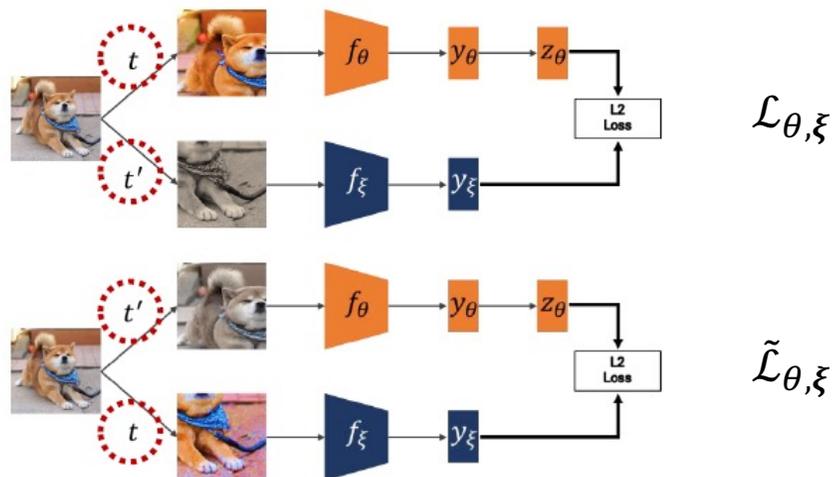
$\overline{q_{\theta}(z_{\theta})} \triangleq \frac{q_{\theta}(z_{\theta})}{\|q_{\theta}(z_{\theta})\|_2}$

$\overline{z'_{\xi}} \triangleq \frac{z'_{\xi}}{\|z'_{\xi}\|_2}$

- Loss Function symmetrization

- ✓ Augmentation 조합을 교환하여 loss를 한번 더 계산

$$\mathcal{L}_{\theta,\xi}^{BYOL} = \mathcal{L}_{\theta,\xi} + \tilde{\mathcal{L}}_{\theta,\xi}$$



02. Design: Optimization

◆ Optimization strategy of BYOL

- ✓ Online Network와 Target Network의 파라미터는 서로 다른 방식으로 갱신된다.
- ✓ 앞에서 계산한 loss function은 Online Network의 파라미터 θ 에 대해서만 최적화됨
- ✓ Target Network의 파라미터 ξ 의 경우, 앞서 언급했던 것처럼 θ 를 이용한 Exponential moving average를 통해 업데이트됨

$$\theta = \text{optimizer}(\theta, \nabla_{\theta} \mathcal{L}_{\theta, \xi}^{BYOL}, \eta) \quad \eta = \text{learning rate}$$

$$\xi = \tau \xi + (1 - \tau) \theta$$

02. Design: Collapsed representation

◆ Collapsed representation

- Positive pair만 사용하였을 때 Mode Collapse가 발생하는 이유
 - ✓ Contrastive learning의 경우, negative samples을 제외하면 별도의 규제항이 없기 때문에, positive samples에 overfitting되면서 collapsed representation을 출력
 - ✓ collapsed representation을 내보내는 것은 모델의 출력이 바뀌지 않는다는 것을 의미하며, 이는 모델이 local optimum problem에 빠져 있다는 것을 의미함 (optimization을 해줄 때 gradient descent에 대한 local minima 문제 발생)
 - ✓ 즉, positive sample의 표현을 출력하도록 학습하는 것이 아니라, 단순히 loss만을 최소화하도록 학습할 경우 Mode Collapse가 발생함
 - ✓ $\arg \min_{\theta, \xi} \nabla_{\theta, \xi} \mathcal{L}_{\theta, \xi}$ 인 (θ^*, ξ^*) 에 도달하여 $\mathcal{L}_{\theta, \xi} = 0$ 이 되면 다른 정보들은 고려하지 않고 같은 표현만을 출력하게 됨

02. Design: Collapsed representation

◆ Collapsed representation

- Positive pair만 사용하였을 때 Mode Collapse가 발생하는 이유
 - ✓ 그러나 BYOL은 $\arg \min_{\xi} \nabla_{\xi} \mathcal{L}_{\xi}$ 방향으로 ξ 를 업데이트하지 않음
 - ✓ BYOL에서 Target Network의 parameter tuning은 exponential moving average 방식을 통해 이루어짐
 - ✓ 저자들은 이러한 EMA 방식이 local minima에 빠지지 않게 해주며, (θ^*, ξ^*) 가 동시에 optimal point가 되는 $\mathcal{L}_{\theta, \xi}$ 는 존재하지 않는다고 함
 - ✓ 또한 만약 Target Network의 파라미터 ξ 에 대해 EMA 방식 대신 gradient descent를 적용하면, Mode collapse가 발생한다고 함

03. Evaluation

03. Evaluation

◆ Implement Details

- Datasets
 - ✓ ImageNet ILSVRC-2012 dataset
- Image Augmentations
 - ✓ 이미지 무작위 패치 선택 -> horizontal flip 무작위 적용 -> 224 x 224 크기 resize
 - ✓ Color distortion 적용
- Architecture
 - ✓ Encoder baseline: Resnet-50 사용
 - ✓ Projection, Predictor: MLP 사용
 - ✓ MLP는 linear-batchnorm-relu-linear 순으로 구성
- Optimization
 - ✓ Lars 사용
 - ✓ 1000 epoch동안 재시작 없이 진행

03. Evaluation

◆ Experiment Results

- Linear evaluation on ImageNet

Method	Top-1	Top-5
Local Agg.	60.2	-
PIRL [35]	63.6	-
CPC v2 [32]	63.8	85.3
CMC [11]	66.2	87.0
SimCLR [8]	69.3	89.0
MoCo v2 [37]	71.1	-
InfoMin Aug. [12]	73.0	91.1
BYOL (ours)	74.3	91.6

(a) ResNet-50 encoder.

Method	Architecture	Param.	Top-1	Top-5
SimCLR [8]	ResNet-50 (2×)	94M	74.2	92.0
CMC [11]	ResNet-50 (2×)	94M	70.6	89.7
BYOL (ours)	ResNet-50 (2×)	94M	77.4	93.6
CPC v2 [32]	ResNet-161	305M	71.5	90.1
MoCo [9]	ResNet-50 (4×)	375M	68.6	-
SimCLR [8]	ResNet-50 (4×)	375M	76.5	93.2
BYOL (ours)	ResNet-50 (4×)	375M	78.6	94.2
BYOL (ours)	ResNet-200 (2×)	250M	79.6	94.8

(b) Other ResNet encoder architectures.

Table 1: Top-1 and top-5 accuracies (in %) under linear evaluation on ImageNet.

- Semi-supervised training on ImageNet

Method	Top-1		Top-5	
	1%	10%	1%	10%
Supervised [77]	25.4	56.4	48.4	80.4
InstDisc	-	-	39.2	77.4
PIRL [35]	-	-	57.2	83.8
SimCLR [8]	48.3	65.6	75.5	87.8
BYOL (ours)	53.2	68.8	78.4	89.0

(a) ResNet-50 encoder.

Method	Architecture	Param.	Top-1		Top-5	
			1%	10%	1%	10%
CPC v2 [32]	ResNet-161	305M	-	-	77.9	91.2
SimCLR [8]	ResNet-50 (2×)	94M	58.5	71.7	83.0	91.2
BYOL (ours)	ResNet-50 (2×)	94M	62.2	73.5	84.1	91.7
SimCLR [8]	ResNet-50 (4×)	375M	63.0	74.4	85.8	92.6
BYOL (ours)	ResNet-50 (4×)	375M	69.1	75.7	87.9	92.5
BYOL (ours)	ResNet-200 (2×)	250M	71.2	77.7	89.5	93.7

(b) Other ResNet encoder architectures.

Table 2: Semi-supervised training with a fraction of ImageNet labels.

03. Evaluation

◆ Experiment Results

- Transfer to other classification tasks

Method	Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
<i>Linear evaluation:</i>												
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
<i>Fine-tuned:</i>												
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.

- Transfer to other vision tasks

Method	AP ₅₀	mIoU	Method	pct. < 1.25	Higher better pct. < 1.25 ²	Lower better pct. < 1.25 ³	rms	rel
Supervised-IN [9]	74.4	74.4	Supervised-IN [83]	81.1	95.3	98.8	0.573	0.127
MoCo [9]	74.9	72.5	SimCLR (repro)	83.3	96.5	99.1	0.557	0.134
SimCLR (repro)	75.2	75.2	BYOL (ours)	84.6	96.7	99.1	0.541	0.129
BYOL (ours)	77.5	76.3						

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

(b) Transfer results on NYU v2 depth estimation.

Table 4: Results on transferring BYOL's representation to other vision tasks.

03. Evaluation

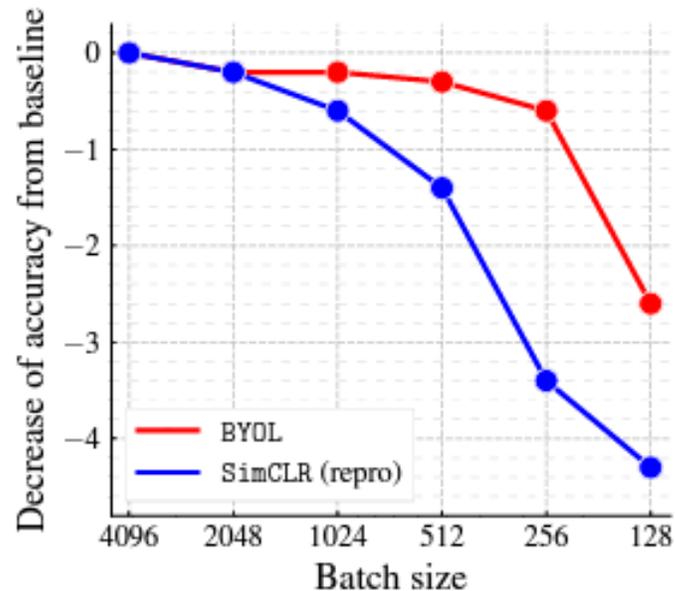
◆ Ablation study

• Batch size

- ✓ 배치 사이즈가 모델의 성능에 미치는 영향을 파악하기 위한 실험
- ✓ SimCLR은 배치 사이즈가 작아짐에 따라서 성능저하가 BYOL보다 가파른 특징을 보임
- ✓ BYOL은 negative sample을 쓰지 않기 때문에 배치사이즈에 강건한 특징을 보임

Batch size	Top-1		Top-5	
	BYOL (ours)	SimCLR (repro)	BYOL (ours)	SimCLR (repro)
4096	72.5	67.9	90.8	88.5
2048	72.4	67.8	90.7	88.5
1024	72.2	67.4	90.7	88.1
512	72.2	66.5	90.8	87.6
256	71.8	64.3±2.1	90.7	86.3±1.0
128	69.6±0.5	63.6	89.6	85.9
64	59.7±1.5	59.2±2.9	83.2±1.2	83.0±1.9

Table 16: Influence of the batch size.



(a) Impact of batch size

03. Evaluation

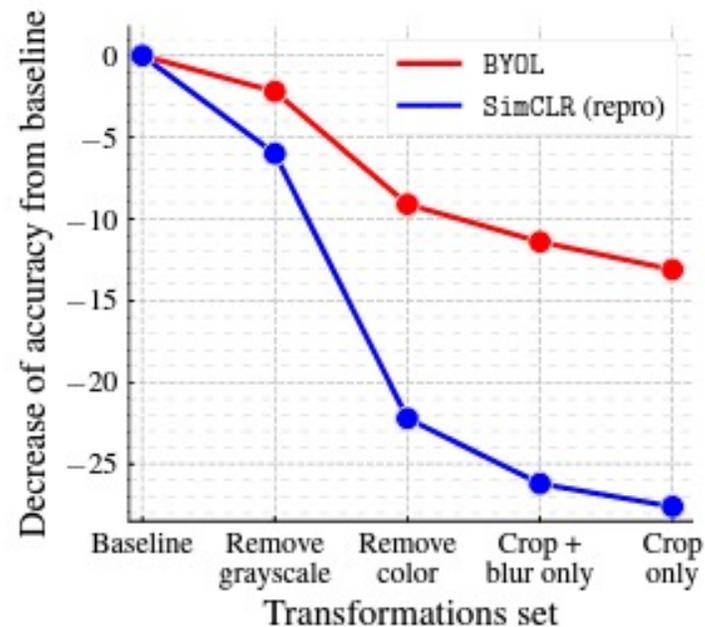
◆ Ablation study

• Image Augmentation

- ✓ Ablation study에서 BYOL과 SimCLR 모두 color distortion을 data augmentation에서 제외했을 때 성능 하락이 크게 나타남
- ✓ BYOL의 경우 SimCLR에 비해 Image Augmentation에서 비교적 강건한 성능을 보여줌

Image augmentation	Top-1		Top-5	
	BYOL (ours)	SimCLR (repro)	BYOL (ours)	SimCLR (repro)
Baseline	72.5	67.9	90.8	88.5
Remove flip	71.9	67.3	90.6	88.2
Remove blur	71.2	65.2	90.3	86.6
Remove color (jittering and grayscale)	63.4±0.7	45.7	85.3±0.5	70.6
Remove color jittering	71.8	63.7	90.7	85.9
Remove grayscale	70.3	61.9	89.8	84.1
Remove blur in \mathcal{T}'	72.4	67.5	90.8	88.4
Remove solarize in \mathcal{T}'	72.3	67.7	90.8	88.2
Remove blur and solarize in \mathcal{T}'	72.2	67.4	90.8	88.1
Symmetric blurring/solarization	72.5	68.1	90.8	88.4
Crop only	59.4±0.3	40.3±0.3	82.4	64.8±0.4
Crop and flip only	60.1±0.3	40.2	83.0±0.3	64.8
Crop and color only	70.7	64.2	90.0	86.2
Crop and blur only	61.1±0.3	41.7	83.9	66.4

Table 17: Ablation on image transformations.



03. Evaluation

◆ Ablation study

• Bootstapping

- ✓ Ablation study에서 exponential moving average coefficient에 대해 실험하였음
- ✓ τ_{base} 가 0일 경우 Target Network에서 Online Network의 파라미터를 그대로 가져오는 것이며, 학습이 되지 않는 모습을 보임
- ✓ τ_{base} 가 1일 경우 Target Network의 파라미터가 Update되지 않는 것이며, 낮은 score를 보이고 있음
- ✓ 가장 학습이 잘되는 τ_{base} 의 값은 0.99이다.

Target	τ_{base}	Top-1
Constant random network	1	18.8 \pm 0.7
Moving average of online	0.999	69.8
Moving average of online	0.99	72.5
Moving average of online	0.9	68.4
Stop gradient of online [†]	0	0.3

(a) Results for different target modes. [†]In the *stop gradient of online*, $\tau = \tau_{base} = 0$ is kept constant throughout training.

04. Conclusion

04. Conclusion

- ◆ Negative sample에 의존하지 않고도 representation learning을 할 수 있는 방법을 찾아내었다.
- ◆ Batch size, augmentation methods의 변화에 대해서도 contrastive methods보다 robust한 모습을 보인다.
- ◆ BYOL을 통해 여러 Task들에 대해 SOTA 달성
- ◆ BYOL은 비전 분야에만 국한된 모델이기 때문에, 다른 분야에 적용하기 위해서는 더 연구가 필요할 수 있다.

Thank You
감사합니다