BYOL

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

▼ Contrastive learning

- No decoder
- No proxy task
- With InfoNCE loss
  - Positive sample간의 거리는 가까이
  - Negative sample과의 거리는 멀리

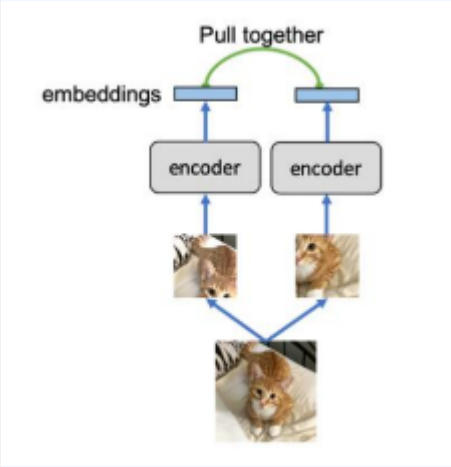
Limitation

1. InfoNCE를 이용하다보니, negative sample의 수에 민감함.
2. 같은 Class의 Sample도 밀어내도록 학습이 진행.
3. Augmentation 조합에 민감함.

Collapsing Problem 해결

Collapsing Problem 이란?

- Positive sample만을 이용하면, 출력되는 두 embeddings을 같게 만드는 supervision만 적용.
- 모델입장에서는 항상 상수만 출력해도 두 embeddings간의 거리가 0이 되어, shortcut을 이용한 학습이 가능함.
- 따라서 이 Collapsing problem을 해결하는 것이 positive-only method의 관건



method

1. Online network, Target network

   서로 다른 네트워크, 서로 다른 입력 영상

   Online network의 Prediction head를 통해서 Target network의 projection 값을 예측.

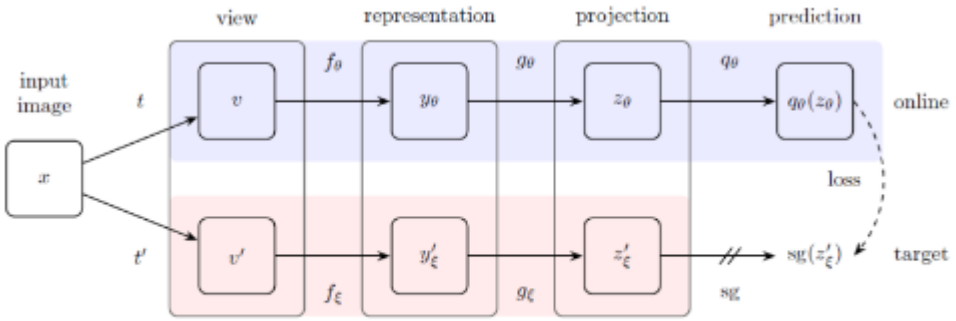


Figure 2: BYOL's architecture. BYOL minimizes a similarity loss between $q_\theta(z_\theta)$ and $\mathrm{sg}(z'_\xi)$, where θ are the trained weights, ξ are an exponential moving average of θ and sg means stop-gradient. At the end of training, everything but f_θ is discarded, and y_θ is used as the image representation.

Online network입장에서는 어떤 view가 Target network에 입력이 될 지 알 수 없음

Prediction head는 모든 view의 mean값을 예측하도록 학습.

Projection값은 항상 다를 테니, Collapse가 일어나지 않음.

2. L2 loss (not InfoNCE)

   negative sample간의 관계를 생각하지 않아도 되기 때문에, 단순히 negative cosine similarity를 사용.

   Symmetric loss 사용

3. Momentum update & stop gradient

   Target network는 학습을 하지 않고, online network를 통해 momentum update로 진행.

   target projection에 consistency를 적용하는 효과. (MoCo와 유사)

---

evaluation

1. Linear evaluation

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.

2. semi-supervised learning

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.

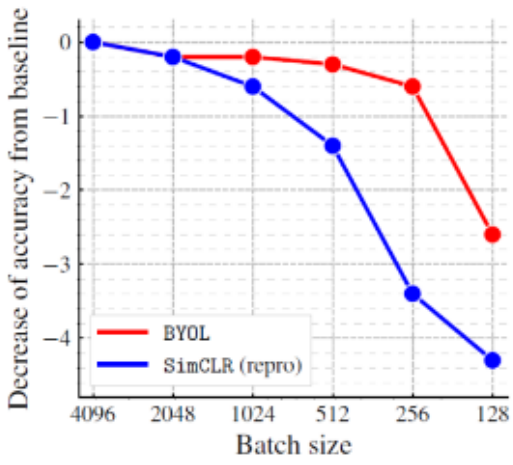
Table 2: Semi-supervised training with a fraction of ImageNet labels.

3. transfer learning

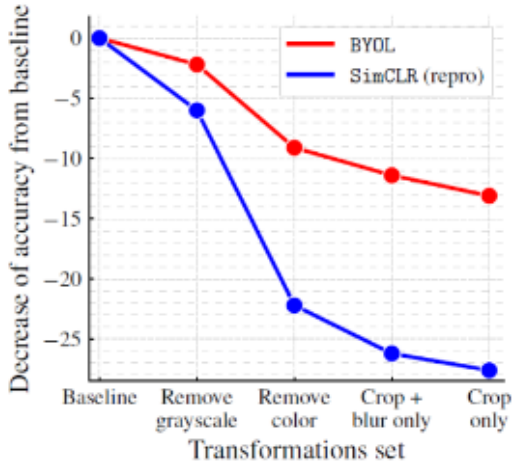
Method	Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
BYOL (ours)	**75.3**	91.3	**78.4**	**57.2**	**62.2**	**67.8**	60.6	82.5	75.5	90.4	94.2	**96.1**
SimCLR (repro)	72.8	90.5	74.4	42.4	60.6	49.3	49.8	81.4	**75.7**	84.6	89.3	92.6
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.

성능이 simclr moco보다 좋은 성능을 보여줌



(a) Impact of batch size



(b) Impact of progressively removing transformations

1. Robustness to batch size & augmentation

   BYOL은 Batch size가 감소하더라도 성능 하락이 크지 않음.

Transformation set에 민감함을 보이지만, simCLR에 비해서는 덜 민감한 성과를 보임.

평가

- BYOL은 Positive sample만을 이용하여 학습한 첫 논문.
- Momentum update와 distillation-based approach를 통해서 collapse problem을 해결.
- 높은 성능과 함께, batch size, augmentation setting에 더 강인함.