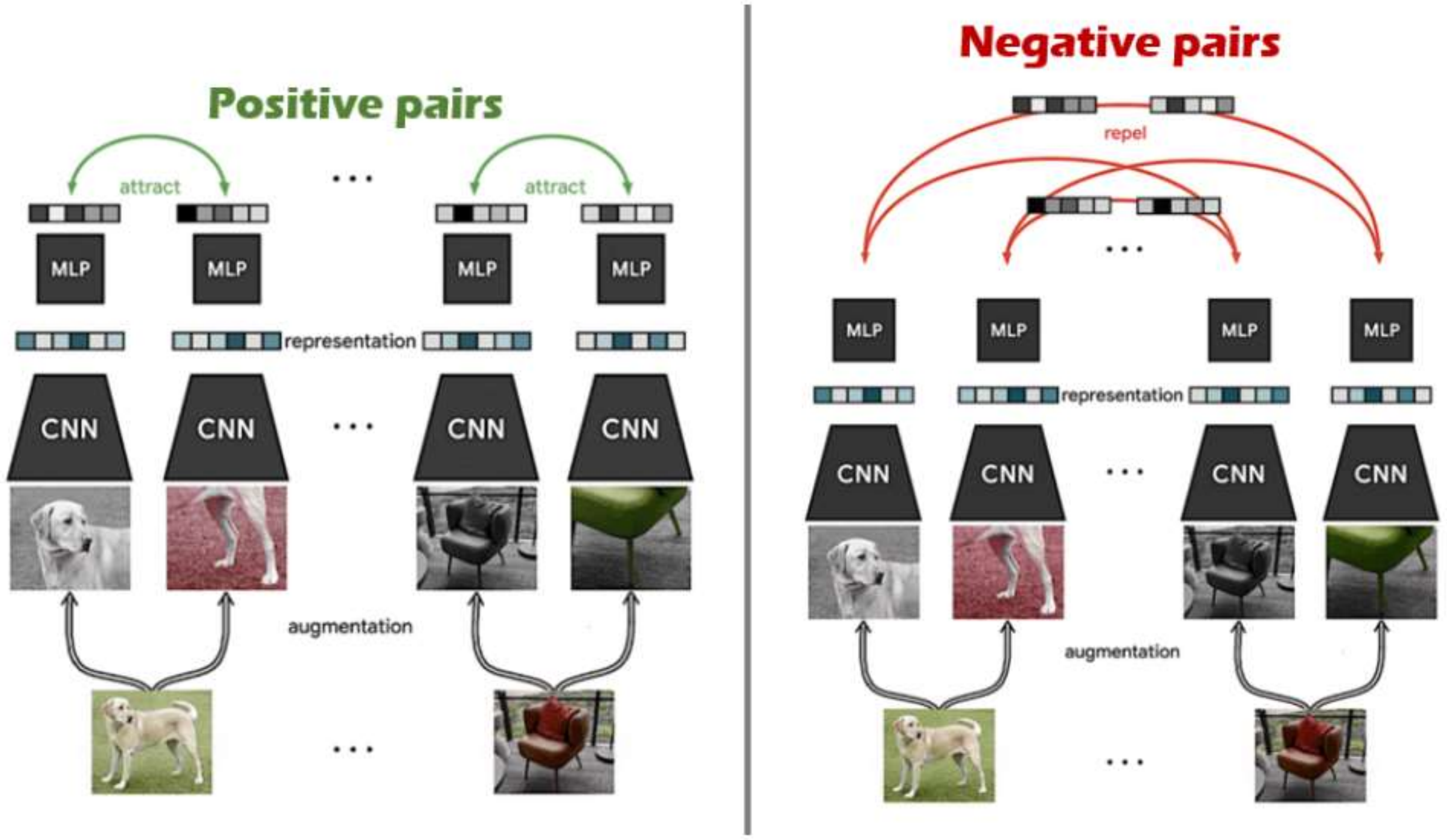
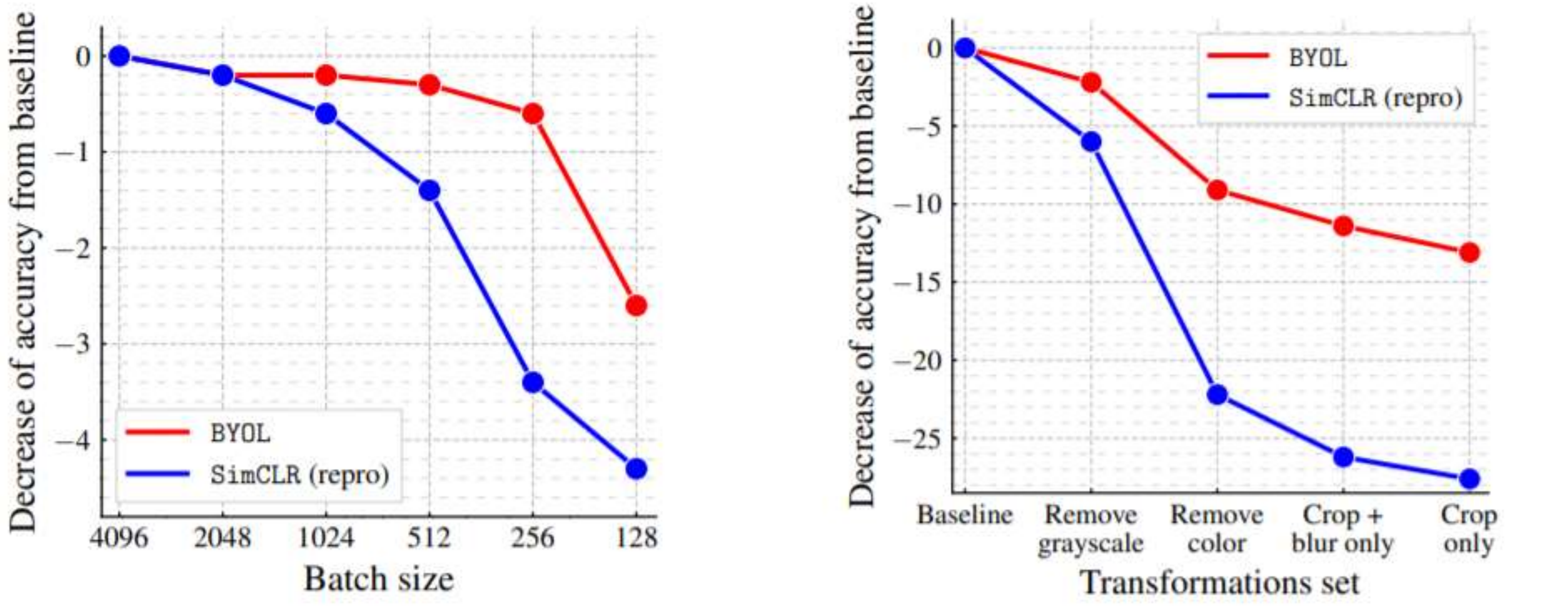
Bootstrap Your Own Latent
A New Approach to
Self-Supervised Learning
(BYOL)

Contrastive learning



-당시 contrastive learning method를 이용하여 feature extractor를 학습시킨 많은 논문이 좋은 성능을 보여주었다.

Contrastive learning- limitations



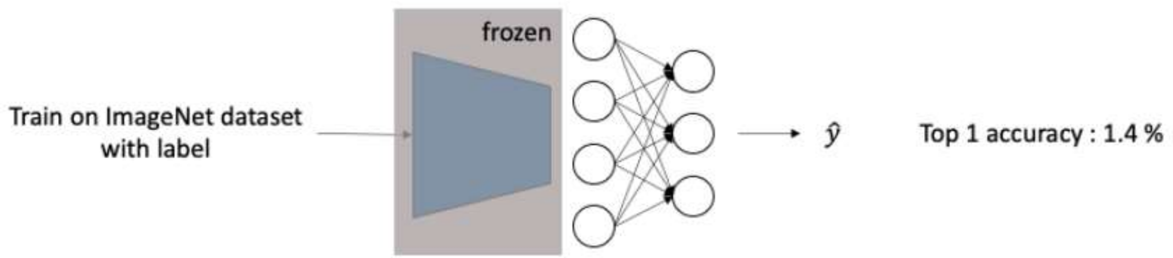
(a) Impact of batch size

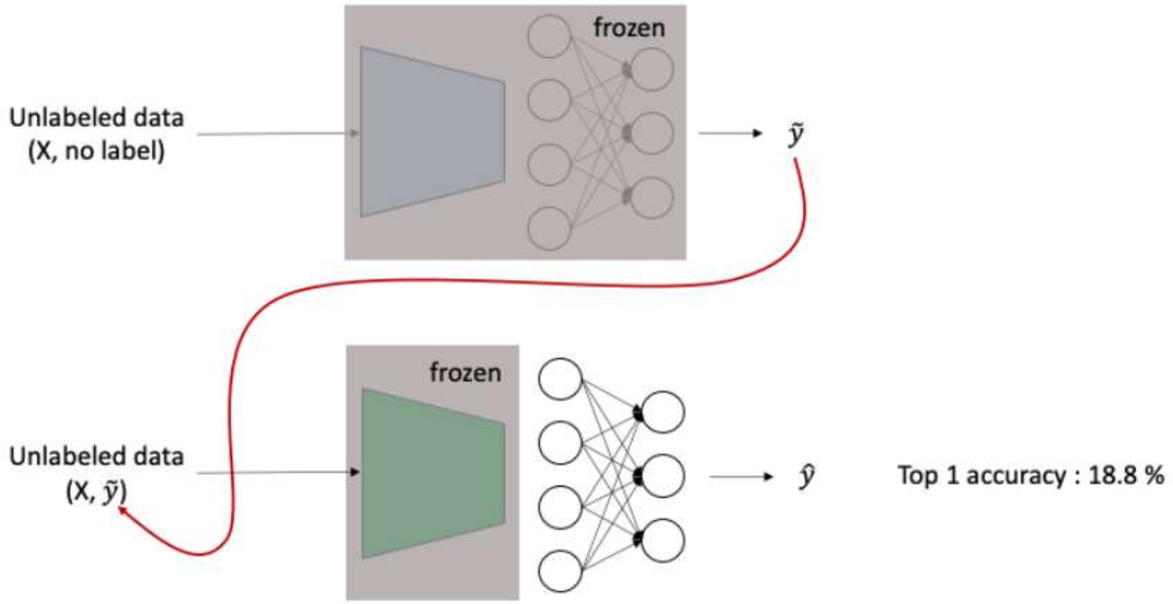
(b) Impact of progressively removing transformations

-negative pairs를 적절히 가져오고 처리해서 사용해야 한다 –
large batch size, memory bank, customized mining strategy 사용, augmentation 방식에 큰 영향

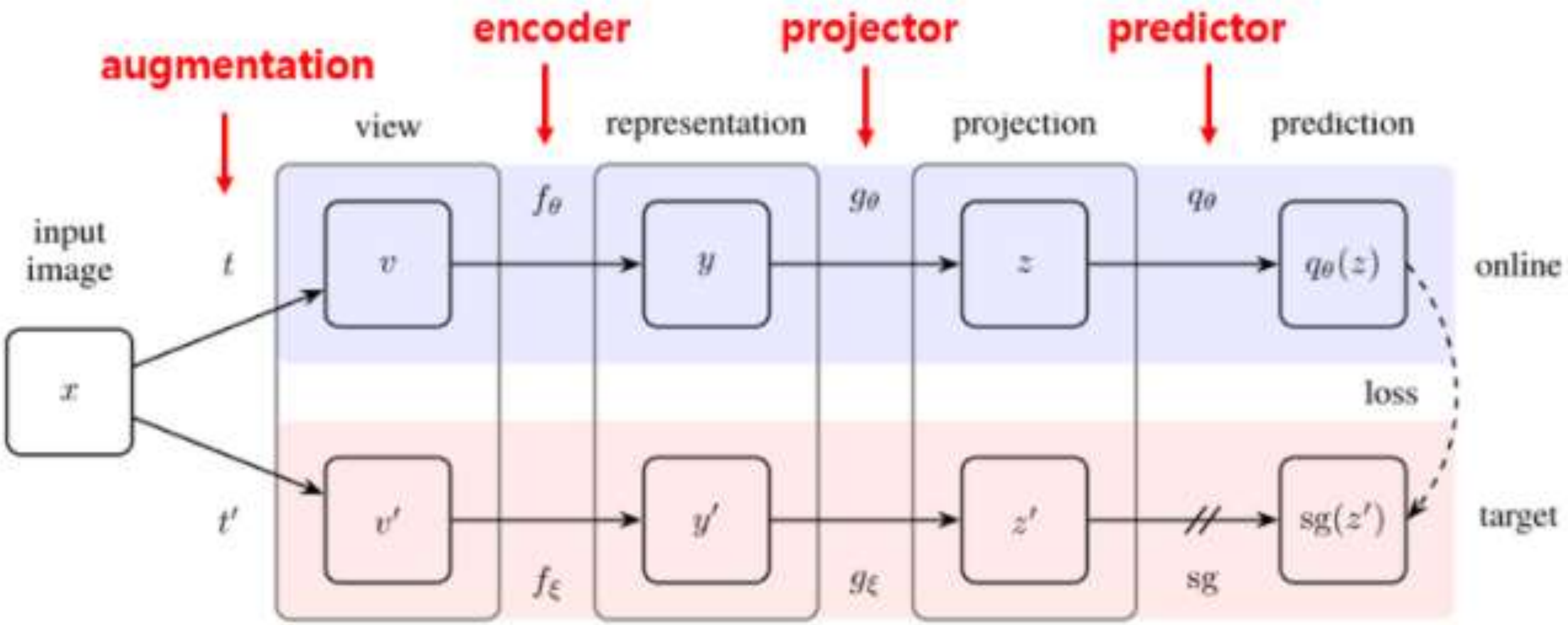
-negative pairs 없이 similarity만 학습할 경우, collapsed representation을 학습할 수 있다.

Idea- Two networks instead of two images





BYOL

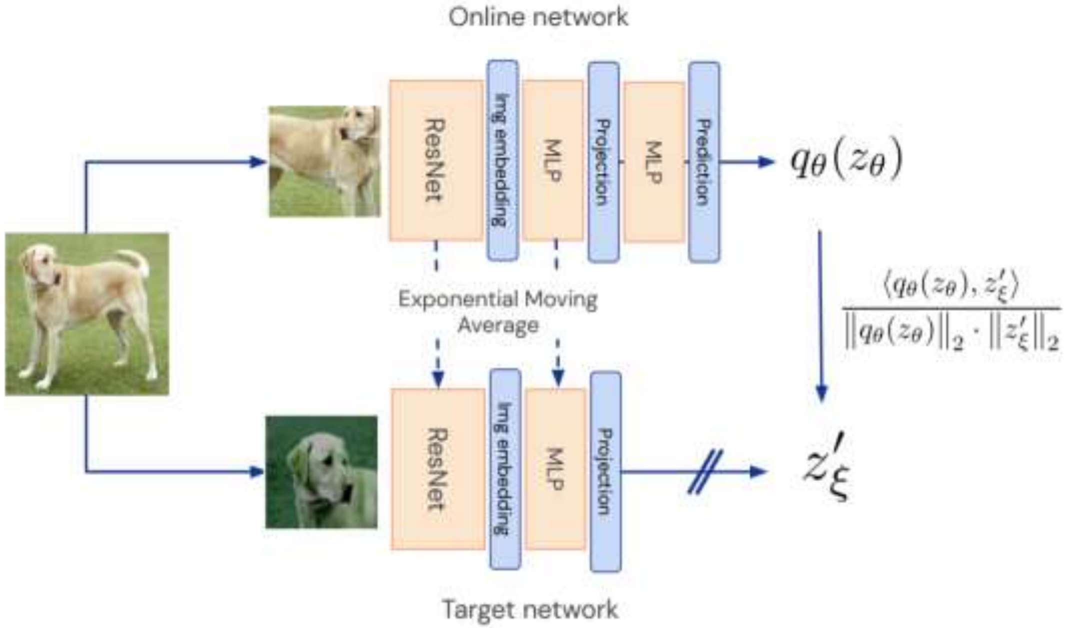


-Downstream task에 사용될 representation을 만들 수 있는 좋은 encoder를 만드는 것이 목적

BYOL- Model

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

$$\theta \leftarrow \text{optimizer}\left(\theta, \nabla_\theta \mathcal{L}^{\texttt{BYOL}}_{\theta,\xi}, \eta\right),$$
$$\xi \leftarrow \tau\xi + (1-\tau)\theta,$$



Exponential moving average, L2 batch normalization, MSE

-Target network의 weight은 loss를 minimum하는 방향으로 나아가지 않고 target과 online의 loss가 같이 loss에 대해 학습하지 않으므로 collapsed representation을 학습하지 않는다.

BYOL- Algorithm

Algorithm 1: BYOL: **Bootstrap Your Own Latent**

Inputs :

$\mathcal{D}$, $\mathcal{T}$, and $\mathcal{T}'$	set of images and distributions of transformations
θ, f_θ, g_θ, and q_θ	initial online parameters, encoder, projector, and predictor
ξ, f_ξ, g_ξ	initial target parameters, target encoder, and target projector
optimizer	optimizer, updates online parameters using the loss gradient
K and N	total number of optimization steps and batch size
$\{\tau_k\}_{k=1}^K$ and $\{\eta_k\}_{k=1}^K$	target network update schedule and learning rate schedule

for $k = 1$ **to** K **do**

$\mathcal{B} \leftarrow \{x_i \sim \mathcal{D}\}_{i=1}^N$ `// sample a batch of` N `images`

for $x_i \in \mathcal{B}$ **do**

$t \sim \mathcal{T}$ and $t' \sim \mathcal{T}'$ `// sample image transformations`

$z_1 \leftarrow g_\theta(f_\theta(t(x_i)))$ and $z_2 \leftarrow g_\theta(f_\theta(t'(x_i)))$ `// compute projections`

loss function symmetrization

$z'_1 \leftarrow g_\xi(f_\xi(t'(x_i)))$ and $z'_2 \leftarrow g_\xi(f_\xi(t(x_i)))$ `// compute target projections`

$l_i \leftarrow -2 \cdot \left(\frac{\langle q_\theta(z_1), z'_1 \rangle}{\|q_\theta(z_1)\|_2 \cdot \|z'_1\|_2} + \frac{\langle q_\theta(z_2), z'_2 \rangle}{\|q_\theta(z_2)\|_2 \cdot \|z'_2\|_2} \right)$ `// compute the loss for` x_i

end

$\delta\theta \leftarrow \frac{1}{N} \sum_{i=1}^{N} \partial_\theta l_i$ `// compute the total loss gradient w.r.t.` θ

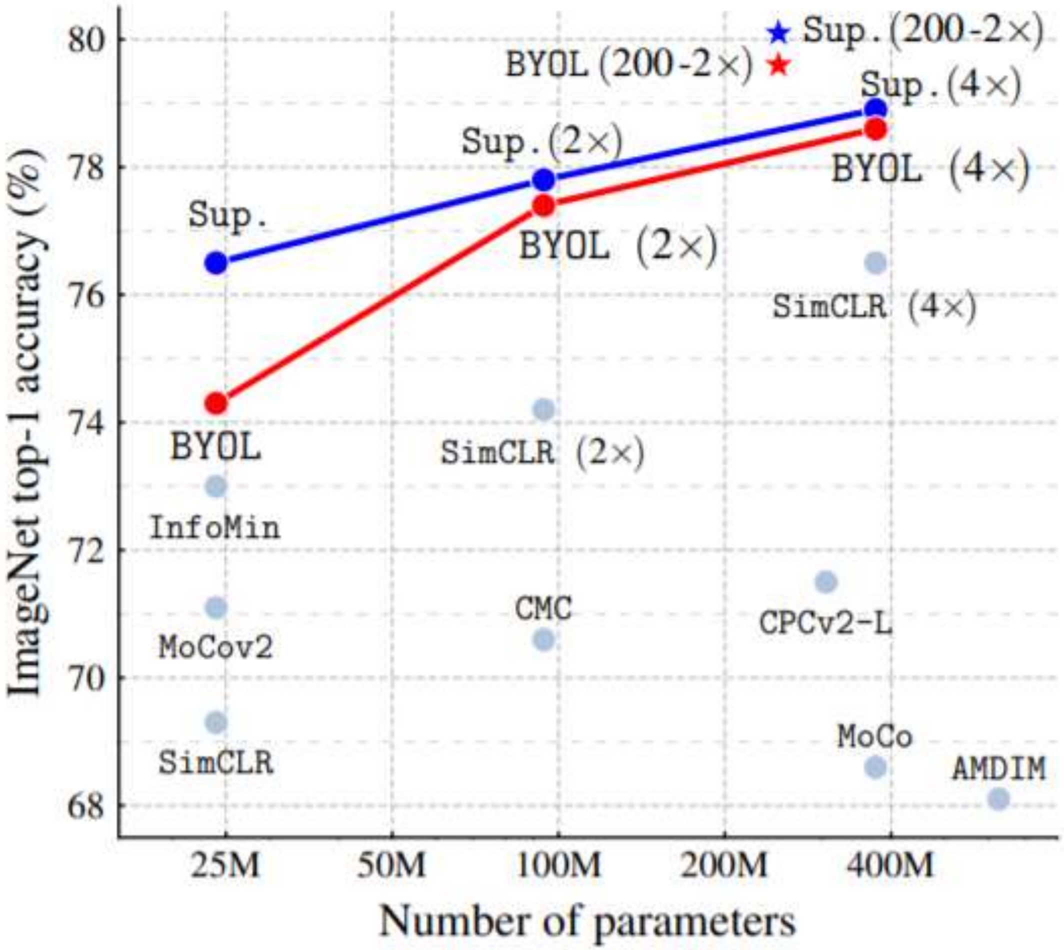
$\theta \leftarrow \text{optimizer}(\theta, \delta\theta, \eta_k)$ `// update online parameters`

$\xi \leftarrow \tau_k \xi + (1 - \tau_k)\theta$ `// update target parameters`

end

Output : encoder f_θ

BYOL- Performance



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.

BYOL- Performance

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.

-ImageNet에서 학습시킨 encode를 다른 data set에 Transfer learning 했을 때 많은 항목에서 Supervised model에 비해서도 좋은 성능을 보였다.

BYOL- Predictor

Method	Predictor	Target network	β	Top-1
BYOL	✓	✓	0	**72.5**
—	✓	✓	1	70.9
—		✓	1	70.7
SimCLR			1	69.4
—	✓		1	69.1
—	✓		0	0.3
Mean Teacher -> —		✓	0	0.2
—			0	0.1

(b) Intermediate variants between BYOL and SimCLR.