Momentum Contrast for Unsupervised Visual Representation Learning

2^{nd} Paper Study | 2022.07.20 Wed.

한 다 희 Han Dahee

Outline

Abstract

✓ Momentum Contrast (MoCo) for unsupervised visual representation learning

✓ A dynamic dictionary with queue and a moving-averaged encoder

→ Large & consistent dictionary

→ Contrastive unsupervised learning을 용이하게

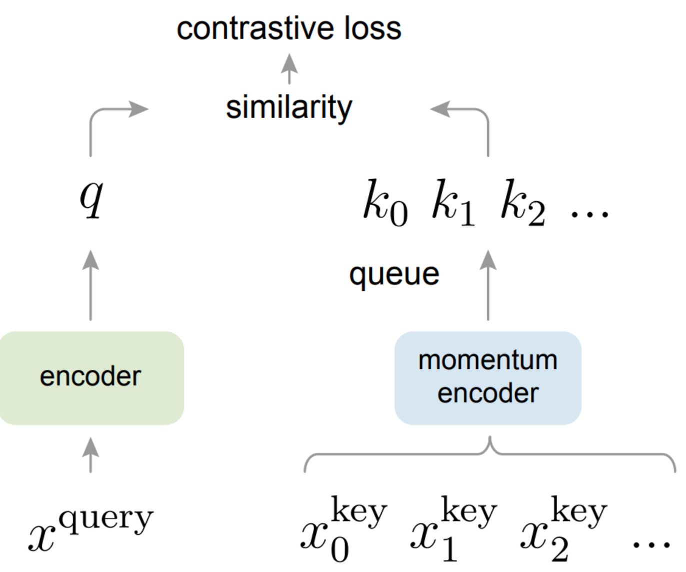
✓ Representation → downstream task good

✓ Supervised pre-training in 7 tasks on PASCAL VOC, COCO, and others와 비교

Introduction

- ✓ Unsupervised representation learning은 자연어처리 분야에서 성공적인 연구
    - ✓ 하지만 여전히 computer vision에서는 supervised learning이 지배적
    - ∵ continuous, high-dimensional space 그리고 구조적이지 않은 데이터 → dictionary building이 쉽지 않음
- ✓ 최근 연구에서 contrastive loss 접근법을 활용한 unsupervised visual representation이 좋은 성과
- ✓ 1. Large : 다양한 negative pair 확보→ 좋은 feature
- ✓ 2. Consistent : 일관된 표현을 위해 느리게 update 되는 key encoder
- ✓ *"MoCo는 contrastive learning을 위해 dynamic dictionary를 building하는 메커니즘, 다양한 pretext task에 사용됨"*

Introduction



← Samples, keys

Positive Sample : 1개
Negative Sample : 나머지

Figure 1. Momentum Contrast (MoCo) trains a visual representation encoder by matching an encoded query q to a dictionary of encoded keys using a contrastive loss. The dictionary keys $\{k_0, k_1, k_2, ...\}$ are defined on-the-fly by a set of data samples. The dictionary is built as a queue, with the current mini-batch enqueued and the oldest mini-batch dequeued, decoupling it from the mini-batch size. The keys are encoded by a slowly progressing encoder, driven by a momentum update with the query encoder. This method enables a large and consistent dictionary for learning visual representations.

Method

- ✓ Contrastive Learning as Dictionary Look-up
  - ✓ Dictionary look-up task를 위한 encoder를 학습시키는 것
  - ✓ k_+ : q에 match 되는 single key
  - ✓ $\{k_0, k_1, k_2 \dots\}$: 나머지 negative samples
  - ✓ τ : hyper parameter
  - ✓ InfoNCE

이미지 유사도 측정

$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i / \tau)} \qquad (1)$$

Method

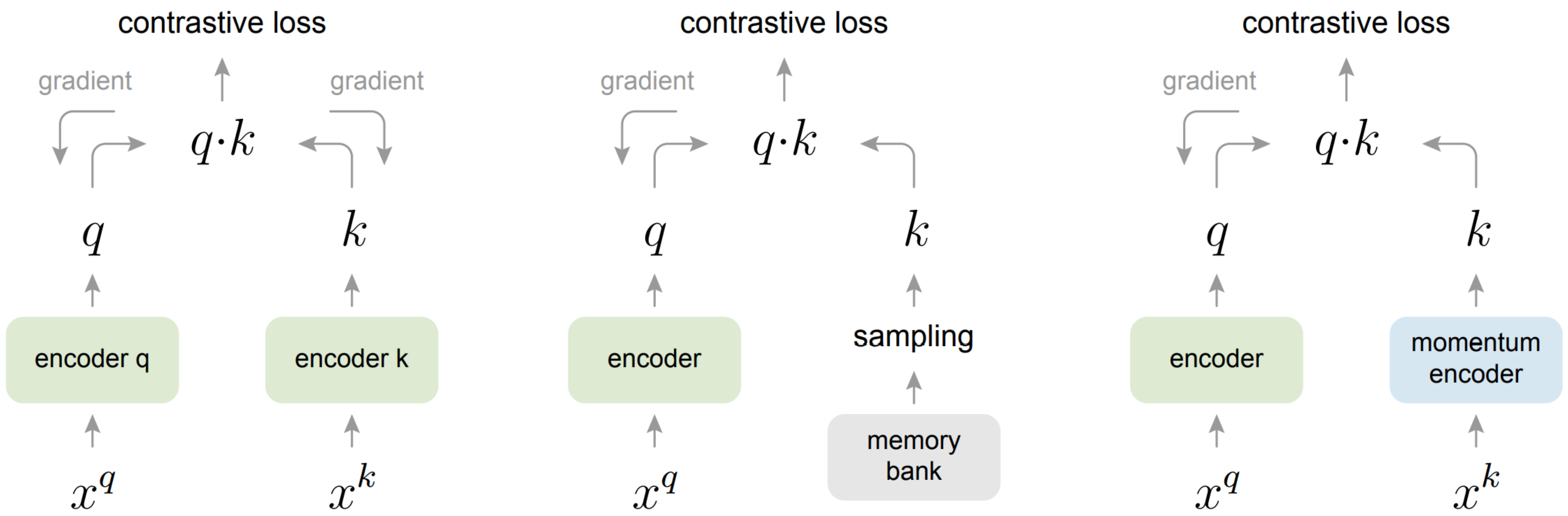
- ✓ Momentum Contrast
  - ✓ Contrastive learning ← 이미지와 같은 high-dimensional continuous input에 대해 discrete dictionary를 만들기위해
  - ✓ Key encoder update → Sampling → “dynamic dictionary”
  - ✓ Dictionary as a queue : current mini-batch to the dictionary, the oldest mini-batch is removed (→ consistency 유지)

Method

- ✓ Momentum Contrast
  - ✓ Momentum update
  - ✓ Large dictionary → queue 사용 → key encoder update 어려움 → encoder copy (q → k)
    → poor result (consistency x) → ***momentum update 제안***
  - ✓ $$\theta_{\mathrm{k}} \leftarrow m\theta_{\mathrm{k}} + (1-m)\theta_{\mathrm{q}}. \qquad (2) \qquad m \in [0, 1)$$
  - ✓ M값이 클수록 better (m = 0.999 default)

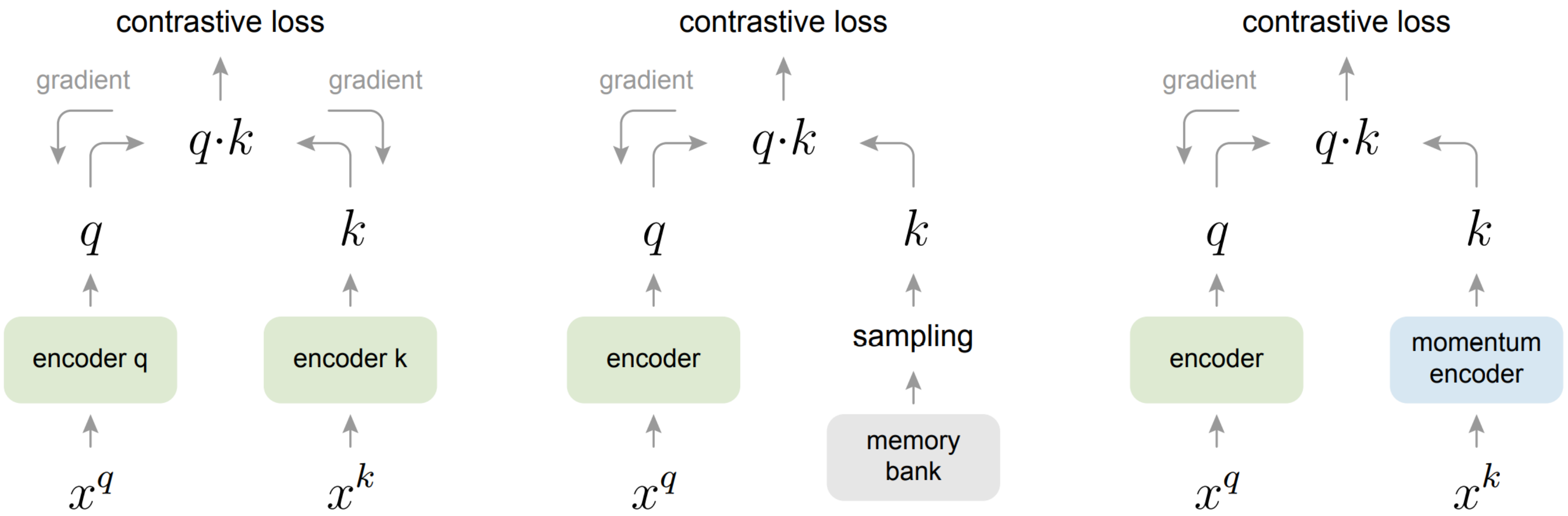
Method

- ✓ Momentum Contrast
  - ✓ Relations to previous mechanisms (dictionary size, consistency)



(a) end-to-end (b) memory bank (c) MoCo

Method



(a) end-to-end (b) memory bank (c) MoCo

- Dictionary = samples = 현재 mini batch
- GPU memory size에 따른 제약
- Key consistency

- Enqueue, dequeue 과정 → memory save
- Key consistency (slowly updated)

- Sample = randomly sampled
- No back-propagation
- Key consistency x

Method

- ✓ Pretext Task
  - ✓ Contrastive learning → 다양한 pretext tasks → “instance discrimination”
  - ✓ Query - k_+ : positive pair
  - ✓ Query - 나머지 : negative pair
- ✓ ResNet as the encoder
  - ✓ Output dimension : 128
  - ✓ L2 norm → representation
  - ✓ τ : 0.07
  - ✓ Data augmentation : color jittering, random horizontal flip, random grayscale conversion …

Method

- ✓ Shuffling BN
  - ✓ BN → good representation 학습 방해
  - ✓ Pretext task cheat 하는 경향, low-loss solution을 쉽게 찾음 ← information leak

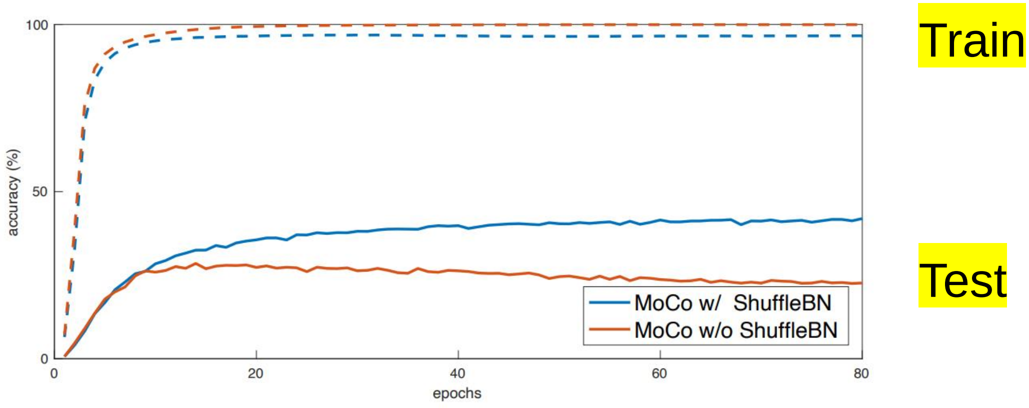
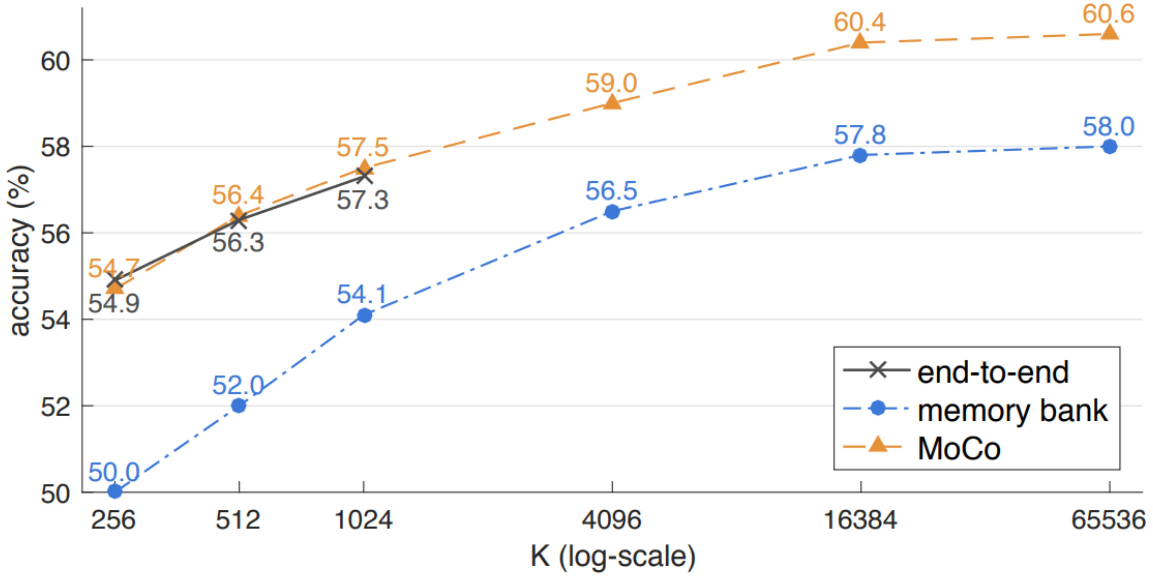


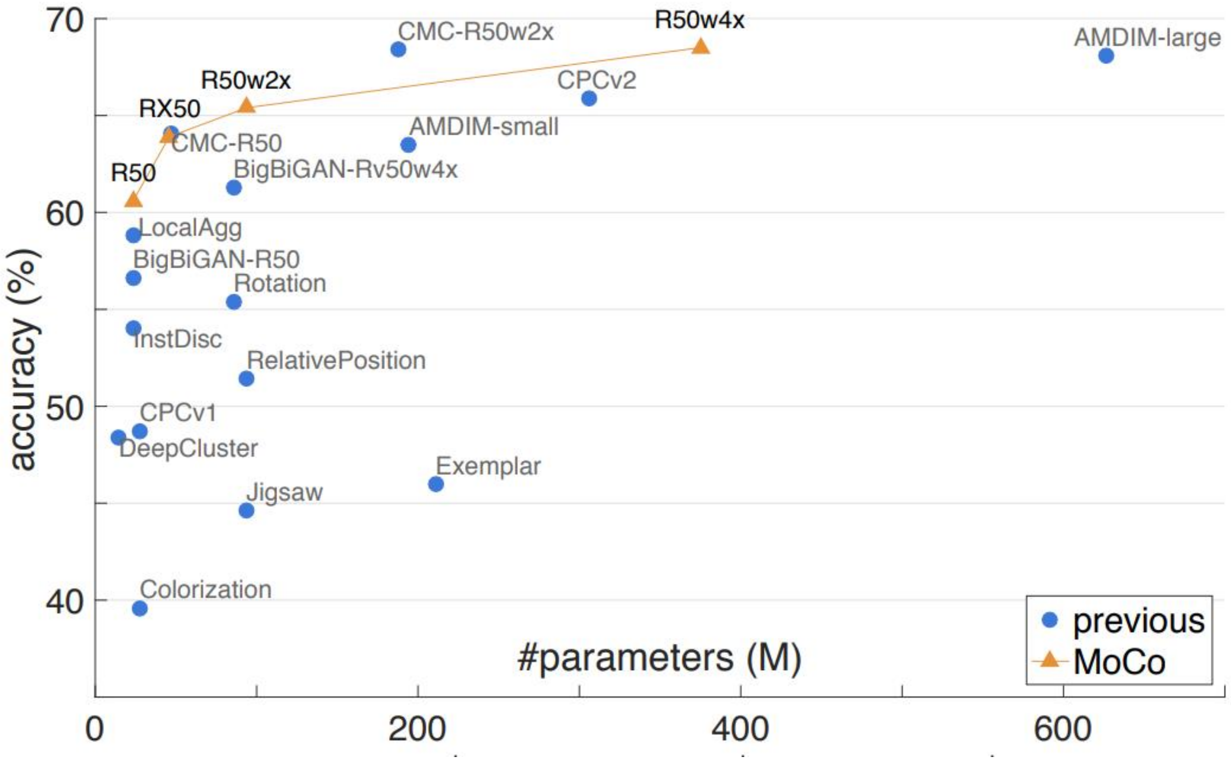
Figure A.1. **Ablation of Shuffling BN**. *Dash*: training curve of the pretext task, plotted as the accuracy of $(K+1)$-way dictionary lookup. *Solid*: validation curve of a kNN-based monitor [61] (not a linear classifier) on ImageNet classification accuracy. This plot shows the first 80 epochs of training: training longer without shuffling BN overfits more.

Experiments



momentum m	0	0.9	0.99	0.999	0.9999
accuracy (%)	*fail*	55.2	57.8	59.0	58.9

Experiments



Comparison under the linear classification protocol on ImageNet

Experiments

pre-train	AP_{50}	AP	AP_{75}
random init.	64.4	37.9	38.6
super. IN-1M	81.4	54.0	59.1
MoCo IN-1M	81.1 (−0.3)	54.6 (**+0.6**)	59.9 (**+0.8**)
MoCo IG-1B	81.6 (+0.2)	55.5 (**+1.5**)	61.2 (**+2.1**)

(a) Faster R-CNN, R50-**dilated-C5**

pre-train	AP_{50}	AP	AP_{75}
random init.	60.2	33.8	33.1
super. IN-1M	81.3	53.5	58.8
MoCo IN-1M	81.5 (+0.2)	55.9 (**+2.4**)	62.6 (**+3.8**)
MoCo IG-1B	82.2 (**+0.9**)	57.2 (**+3.7**)	63.7 (**+4.9**)

(b) Faster R-CNN, R50-**C4**

Table 2. **Object detection fine-tuned on PASCAL VOC** `trainval07+12`. Evaluation is on `test2007`: AP_{50} (default VOC metric), AP (COCO-style), and AP_{75}, averaged over 5 trials. All are fine-tuned for 24k iterations (~23 epochs). In the brackets are the gaps to the ImageNet supervised pre-training counterpart. In green are the gaps of at least **+0.5** point.

	R50-dilated-C5			R50-C4		
pre-train	AP_{50}	AP	AP_{75}	AP_{50}	AP	AP_{75}
end-to-end	79.2	52.0	56.6	80.4	54.6	60.3
memory bank	79.8	52.9	57.9	80.6	54.9	60.6
MoCo	**81.1**	**54.6**	**59.9**	**81.5**	**55.9**	**62.6**

Table 3. **Comparison of three contrastive loss mechanisms** on PASCAL VOC object detection, fine-tuned on `trainval07+12` and evaluated on `test2007` (averages over 5 trials). All models are implemented by us (Figure 3), pre-trained on IN-1M, and fine-tuned using the same settings as in Table 2.

Discussion and Conclusion

- ✓ Positive results of unsupervised learning
- ✓ MoCo가 실용적이고, 다른 pretext task에 대해 유용하기를 기대

MoCo v2

SimCLR

1) Large batch size
2) MLP projection head
3) Strong data augmentation

MoCo v2

SimCLR

1) Large batch size
2) MLP projection head
3) Strong data augmentation

MoCo v2

1) Large batch size
2) MLP projection head
3) Strong data augmentation

MoCo v2

case	unsup. pre-train				ImageNet	VOC detection		
	MLP	aug+	cos	epochs	acc.	AP_{50}	AP	AP_{75}
supervised					76.5	81.3	53.5	58.8
MoCo v1				200	60.6	81.5	55.9	62.6
(a)	✓			200	66.2	82.0	56.4	62.6
(b)		✓		200	63.4	82.2	56.8	63.2
(c)	✓	✓		200	67.3	**82.5**	57.2	63.9
(d)	✓	✓	✓	200	67.5	82.4	57.0	63.6
(e)	✓	✓	✓	**800**	**71.1**	**82.5**	**57.4**	**64.0**

Table 1. **Ablation of MoCo baselines**, evaluated by ResNet-50 for (i) ImageNet linear classification, and (ii) fine-tuning VOC object detection (mean of 5 trials). "**MLP**": with an MLP head; "**aug+**": with extra blur augmentation; "**cos**": cosine learning rate schedule.

τ	0.07	0.1	0.2	0.3	0.4	0.5
w/o MLP	60.6	**60.7**	59.0	58.2	57.2	56.4
w/ MLP	62.9	64.9	**66.2**	65.7	65.0	64.3

MoCo v2

case	MLP	aug+	cos	epochs	batch	ImageNet acc.
	unsup. pre-train					
MoCo v1 [6]				200	256	60.6
SimCLR [2]	✓	✓	✓	200	256	61.9
SimCLR [2]	✓	✓	✓	200	8192	66.6
MoCo v2	✓	✓	✓	200	256	**67.5**
results of ***longer*** *unsupervised training follow:*						
SimCLR [2]	✓	✓	✓	1000	4096	69.3
MoCo v2	✓	✓	✓	800	256	**71.1**

Table 2. **MoCo *vs.* SimCLR**: ImageNet linear classifier accuracy (**ResNet-50, 1-crop 224×224**), trained on features from unsupervised pre-training. "aug+" in SimCLR includes blur and stronger color distortion. SimCLR ablations are from Fig. 9 in [2] (we thank the authors for providing the numerical results).

mechanism	batch	memory / GPU	time / 200-ep.
MoCo	256	**5.0G**	**53 hrs**
end-to-end	256	7.4G	65 hrs
end-to-end	4096	93.0G†	n/a

Table 3. **Memory and time cost** in 8 V100 16G GPUs, implemented in PyTorch. †: based on our estimation.

References

[Paper]

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9729-9738).

Chen, X., Fan, H., Girshick, R., & He, K. (2020). Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

[Review]

https://hongl.tistory.com/127?category=934082

https://deepseow.tistory.com/51

Q & A

Thank you ☺