Momentum Contrast for Unsupervised Visual Representation Learning

2023. 05. 18

MoCo (2020, CVPR)

❖ Momentum Contrast for Unsupervised Visual Representation Learning

- 2020년에 CVPR에 게재되었으며, 2023년 05월 18일 기준 7033회 인용됨
- Dynamic dictionary와 Momentum update를 사용한 contrastive learning 방법론
- CV 분야에서 지도 학습과 비지도 학습으로 표현 학습을 했을 때의 성능의 차이를 줄여주는 방법론

Momentum Contrast for Unsupervised Visual Representation Learning

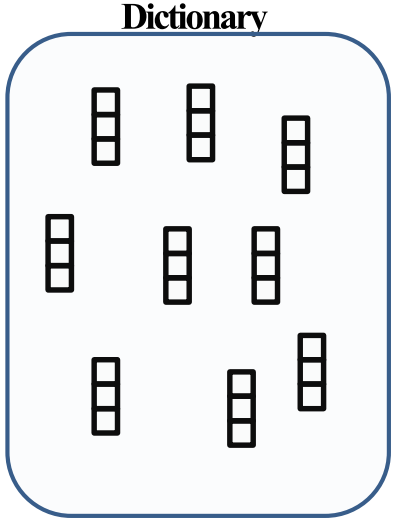
Kaiming He Haoqi Fan Yuxin Wu Saining Xie Ross Girshick

Facebook AI Research (FAIR)

Background

❖ Natural Language Processing (NLP) vs Computer vision (CV)

- **NLP: 비지도 표현 학습을 사용해 레이블을 사용하지 않고도 좋은 표현 학습을 가능하게 한 사례가 있음**
  - ➢ Discrete symbol (ex: word, character…)이 연속으로 구성된 데이터를 사용하기 때문
  - ➢ 자연어 처리에서 사용되는 모델이 텍스트 데이터 자체의 내재적인 구조, 순차적인 특성을 활용해서 좋은 표현 학습 가능
- **CV: 비지도 표현 학습을 사용해 레이블 없이 좋은 표현 학습을 하는 것보다 지도 학습을 사용하는 것이 더 우세**
  - ➢ 이미지 데이터는 연속적이고 고차원 데이터이며 텍스트 데이터처럼 순차적인 특성을 갖고 있지 않음
  - ➢ 따라서, 레이블이 없는 경우 → Pretext task, Contrastive Learning 등장



Proposed Method

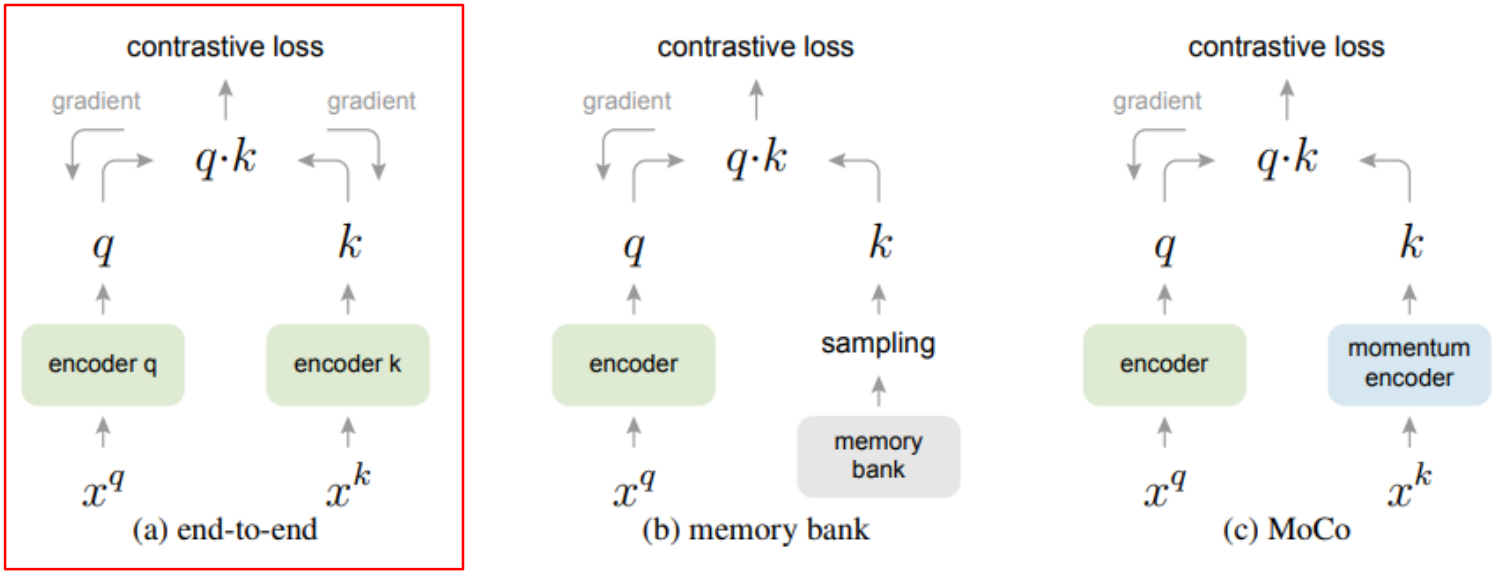
❖ Overall Architecture

- Contrastive loss를 사용한 컨셉이 비슷한 세 가지의 방법론 비교 설명

1. end-to-end

   ✓ 동일한 encoder를 사용해서 query와 key representation vector 추출 → inconsistency

   ✓ Mini batch size 만큼의 dictionary 사용 → dictionary 사이즈가 GPU 메모리 사이즈에 제한 받음



Proposed Method

❖ Overall Architecture

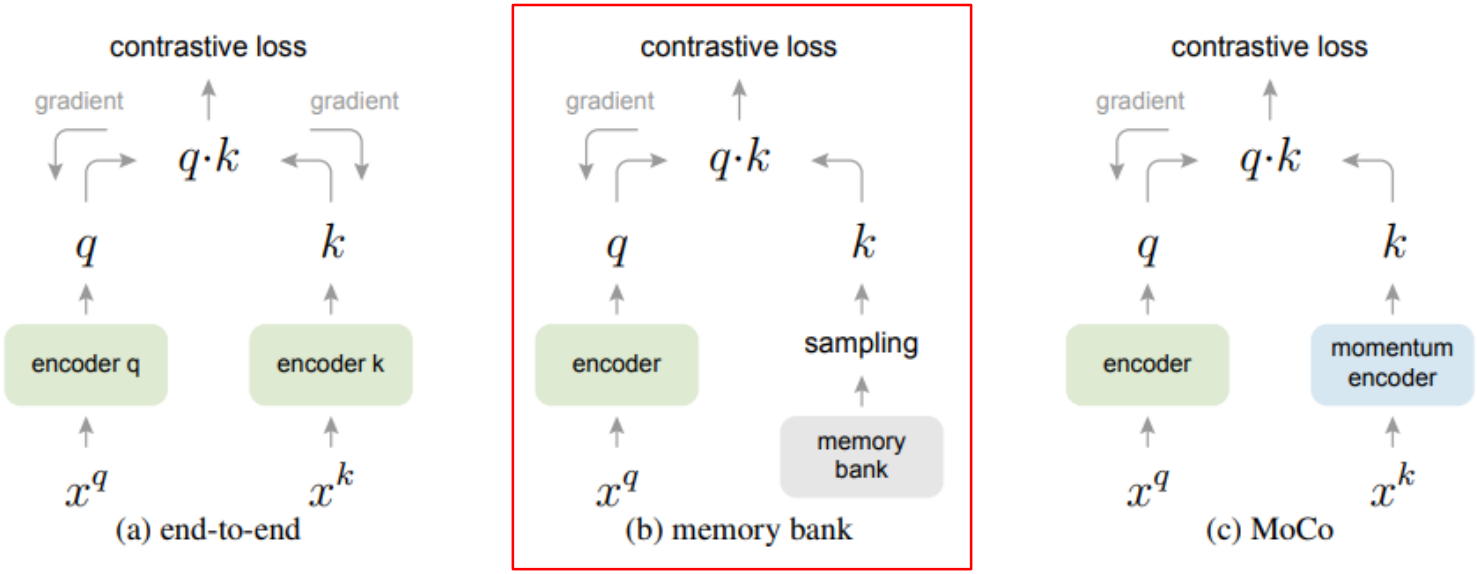
- Contrastive loss를 사용한 컨셉이 비슷한 세 가지의 방법론 비교 설명

1. memory bank

✓ MoCo와 동일하게 momory bank에서 샘플링한 key 값 사용

✓ 모든 데이터 셋에 대한 representation memory bank에 저장

✓ Momentum update를 하는데, encoder에 해당 방법을 적용하는 것이 momory bank 내에 새로운 샘플을 업데이트하는데 적용 → less consistent



Proposed Method

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

❖ Overall Architecture

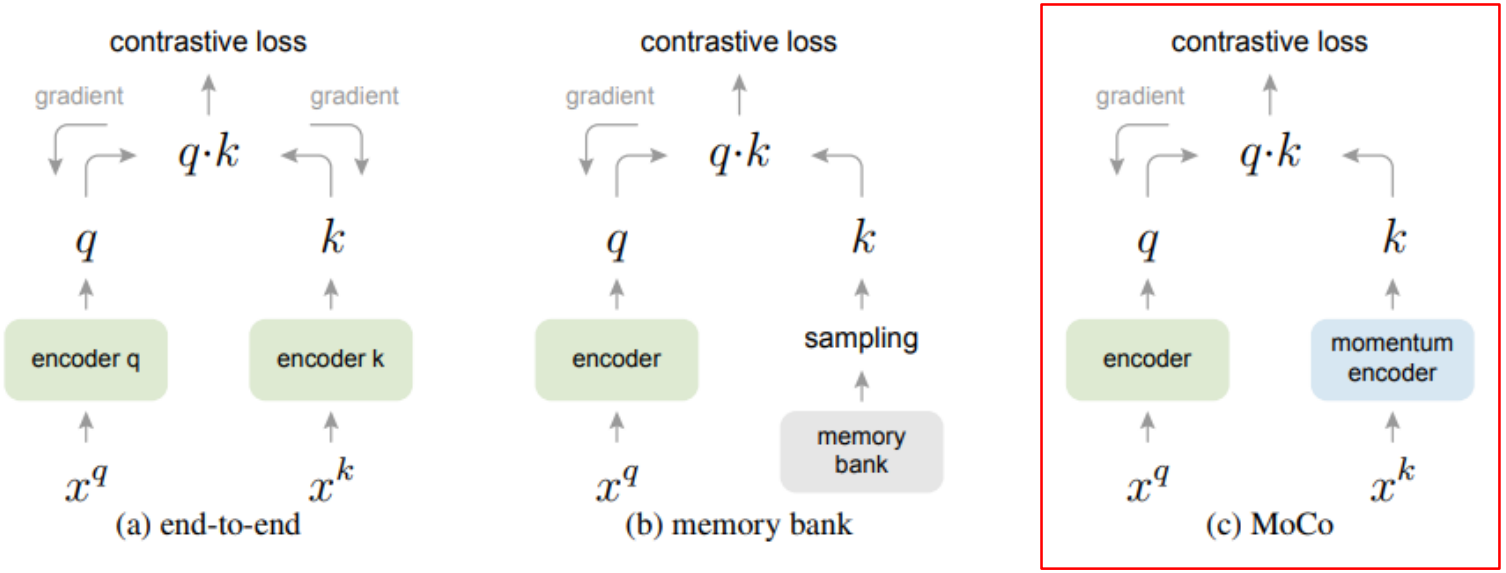
- Contrastive loss를 사용한 컨셉이 비슷한 세 가지의 방법론 비교 설명

1. MoCo

✓ Query encoder에만 gradient update 적용, key encoder에는 momentum update 적용 $\theta_k \leftarrow m\theta_k + (1-m)\theta_q$.

✓ Queue 형태의 dictionary를 사용함으로써 batch size에 영향을 받지 않고 거대한 negative sample 보유 가능

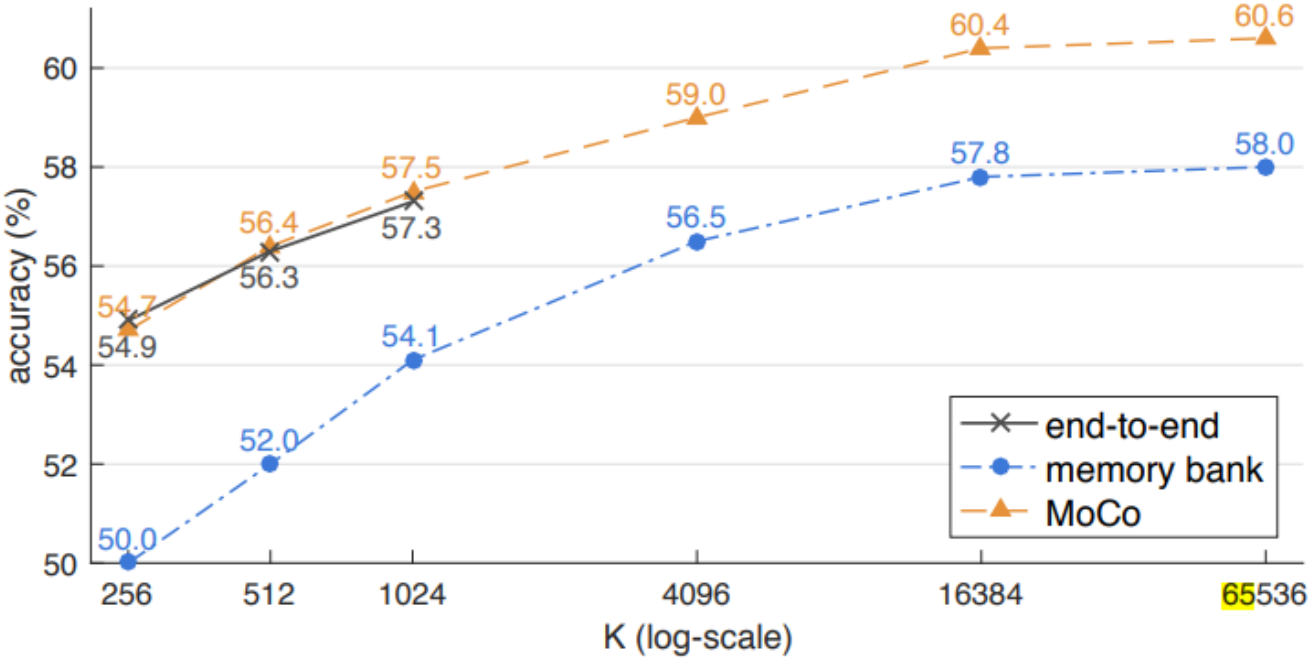
✓ Momentum encoder에서 나온 key가 새로운 sample로 들어가고 오래된 sample 제거



Experiment

❖ Comparison of three contrastive loss mechanisms

- MoCo가 가장 우수한 성능을 보임
- end-to-end의 경우, mini-batch size만큼 dictionary를 갖고 있다 보니, GPU 메모리 제한으로 인해 1024 이후로는 진행 불가능



Experiment

❖ Comparison under the linear classification protocol in ImageNet

method	architecture	#params (M)	accuracy (%)
Exemplar [17]	R50w3×	211	46.0 [38]
RelativePosition [13]	R50w2×	94	51.4 [38]
Jigsaw [45]	R50w2×	94	44.6 [38]
Rotation [19]	Rv50w4×	86	55.4 [38]
Colorization [64]	R101*	28	39.6 [14]
DeepCluster [3]	VGG [53]	15	48.4 [4]
BigBiGAN [16]	R50	24	56.6
	Rv50w4×	86	61.3
methods based on contrastive learning follow:			
InstDisc [61]	R50	24	54.0
LocalAgg [66]	R50	24	58.8
CPC v1 [46]	R101*	28	48.7
CPC v2 [35]	$R170^{*}_{wider}$	303	65.9
CMC [56]	$R50_{L+ab}$	47	$64.1^{\dagger}$
	$R50w2\times_{L+ab}$	188	$68.4^{\dagger}$
AMDIM [2]	$AMDIM_{small}$	194	$63.5^{\dagger}$
	$AMDIM_{large}$	626	$68.1^{\dagger}$
MoCo	R50	24	60.6
	RX50	46	63.9
	R50w2×	94	65.4
	R50w4×	375	**68.6**

Experiment

❖ Comparison with previous methods on object detection fine-tuned on PASCAL VOC

ImageNet-1M

pre-train	Default VOC metric: AP_{50} RelPos, by [14]	Multi-task [14]	Jigsaw, by [26]	LocalAgg [66]	MoCo	COCO style: AP MoCo	Averaged over 5 trials: AP_{75} Multi-task [14]	MoCo
super. IN-1M	74.2	74.2	70.5	74.6	74.4	42.4	44.3	42.7
unsup. IN-1M	66.8 (−7.4)	70.5 (−3.7)	61.4 (−9.1)	69.1 (−5.5)	74.9 (+0.5)	46.6 (+4.2)	43.9 (−0.4)	50.1 (+7.4)
unsup. IN-14M	-	-	69.2 (−1.3)	-	75.2 (+0.8)	46.9 (+4.5)	-	50.2 (+7.5)
unsup. YFCC-100M	-	-	66.6 (−3.9)	-	74.7 (+0.3)	45.9 (+3.5)	-	49.0 (+6.3)
unsup. IG-1B	-	-	-	-	75.6 (+1.2)	47.6 (+5.2)	-	51.7 (+9.0)

Instagram-1B

❖ Object detection and instance segmentation fine-tuned on COCO

pre-train	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
random init.	31.0	49.5	33.2	28.5	46.8	30.4
super. IN-1M	38.9	59.6	42.7	35.4	56.5	38.1
MoCo IN-1M	38.5 (−0.4)	58.9 (−0.7)	42.0 (−0.7)	35.1 (−0.3)	55.9 (−0.6)	37.7 (−0.4)
MoCo IG-1B	38.9 (0.0)	59.4 (−0.2)	42.3 (−0.4)	35.4 (0.0)	56.5 (0.0)	37.9 (−0.2)

(a) Mask R-CNN, R50-**FPN**, **1**× schedule

pre-train	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
random init.	36.7	56.7	40.0	33.7	53.8	35.9
super. IN-1M	40.6	61.3	44.4	36.8	58.1	39.5
MoCo IN-1M	40.8 (+0.2)	61.6 (+0.3)	44.7 (+0.3)	36.9 (+0.1)	58.4 (+0.3)	39.7 (+0.2)
MoCo IG-1B	41.1 (+0.5)	61.8 (+0.5)	45.1 (+0.7)	37.4 (+0.6)	59.1 (+1.0)	40.2 (+0.7)

(b) Mask R-CNN, R50-**FPN**, **2**× schedule

pre-train	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
random init.	26.4	44.0	27.8	29.3	46.9	30.8
super. IN-1M	38.2	58.2	41.2	33.3	54.7	35.2
MoCo IN-1M	38.5 (+0.3)	58.3 (+0.1)	41.6 (+0.4)	33.6 (+0.3)	54.8 (+0.1)	35.6 (+0.4)
MoCo IG-1B	39.1 (+0.9)	58.7 (+0.5)	42.2 (+1.0)	34.1 (+0.8)	55.4 (+0.7)	36.4 (+1.2)

pre-train	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
random init.	35.6	54.6	38.2	31.4	51.5	33.5
super. IN-1M	40.0	59.9	43.1	34.7	56.5	36.9
MoCo IN-1M	40.7 (+0.7)	60.5 (+0.6)	44.1 (+1.0)	35.4 (+0.7)	57.3 (+0.8)	37.6 (+0.7)
MoCo IG-1B	41.1 (+1.1)	60.7 (+0.8)	44.8 (+1.7)	35.6 (+0.9)	57.4 (+0.9)	38.1 (+1.2)

Experiment

- MoCo vs. ImageNet supervised pre-training, fine-tuned on various tasks

pre-train	COCO keypoint detection AP^{kp}	AP^{kp}_{50}	AP^{kp}_{75}
random init.	65.9	86.5	71.7
super. IN-1M	65.8	86.9	71.9
MoCo IN-1M	66.8 (+1.0)	87.4 (+0.5)	72.5 (+0.6)
MoCo IG-1B	66.9 (+1.1)	87.8 (+0.9)	73.0 (+1.1)

pre-train	COCO dense pose estimation AP^{dp}	AP^{dp}_{50}	AP^{dp}_{75}
random init.	39.4	78.5	35.1
super. IN-1M	48.3	85.6	50.6
MoCo IN-1M	50.1 (+1.8)	86.8 (+1.2)	53.9 (+3.3)
MoCo IG-1B	50.6 (+2.3)	87.0 (+1.4)	54.3 (+3.7)

pre-train	LVIS v0.5 instance segmentation AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
random init.	22.5	34.8	23.8
super. IN-1M†	24.4	37.8	25.8
MoCo IN-1M	24.1 (−0.3)	37.4 (−0.4)	25.5 (−0.3)
MoCo IG-1B	24.9 (+0.5)	38.2 (+0.4)	26.4 (+0.6)

pre-train	Cityscapes instance seg. AP^{mk}	AP^{mk}_{50}	Semantic seg. (mIoU) Cityscapes	VOC
random init.	25.4	51.1	65.3	39.5
super. IN-1M	32.9	59.6	74.6	74.4
MoCo IN-1M	32.3 (−0.6)	59.3 (−0.3)	75.3 (+0.7)	72.5 (−1.9)
MoCo IG-1B	32.9 (0.0)	60.3 (+0.7)	75.5 (+0.9)	73.6 (−0.8)