Momentum contrast for Unsupervised representation learning (MoCo)

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross Girshick

Self-supervised learning

- Random initialization vs. Pre-training
- Target of self-supervision - learning transferable features

Image

CNN Backbone

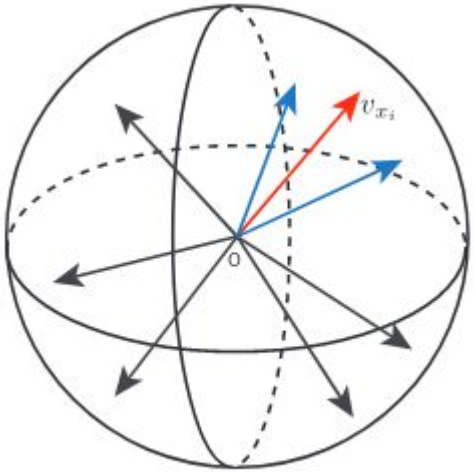
2048D

Low dim embedding

128D

L2 norm

128D

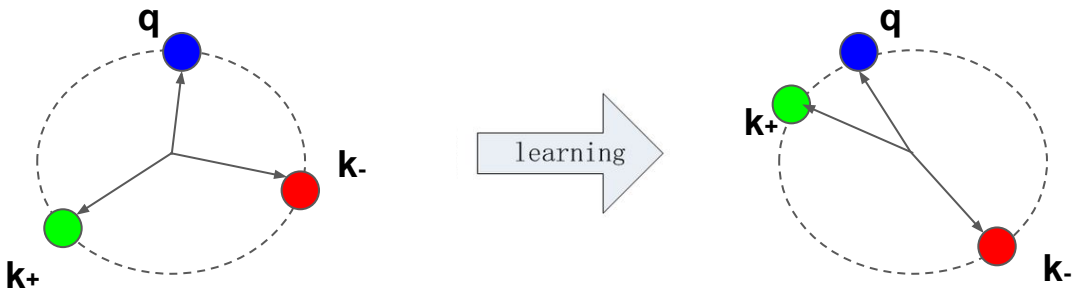


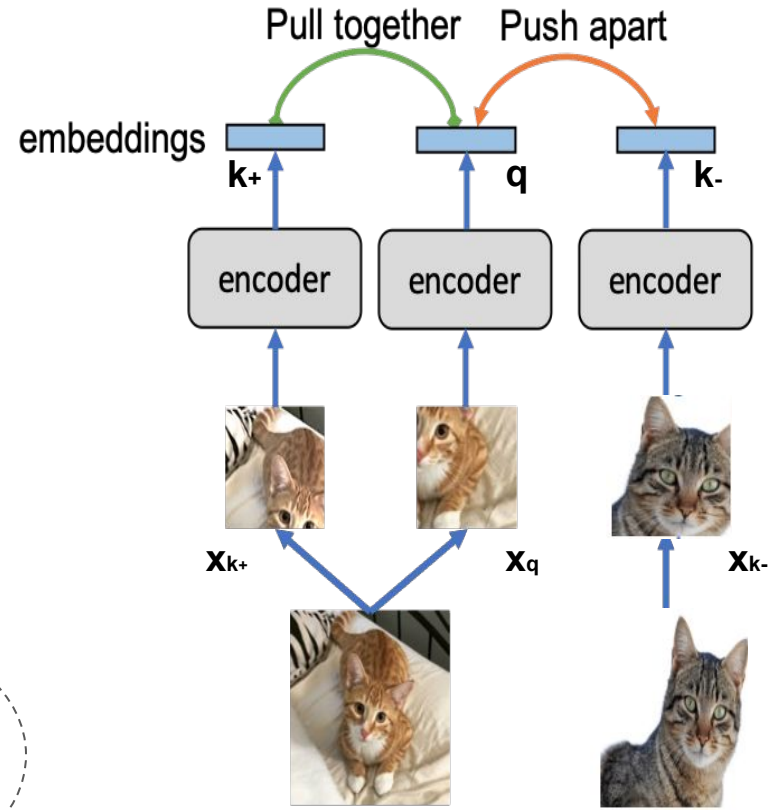
128D Unit Sphere

General Contrastive learning

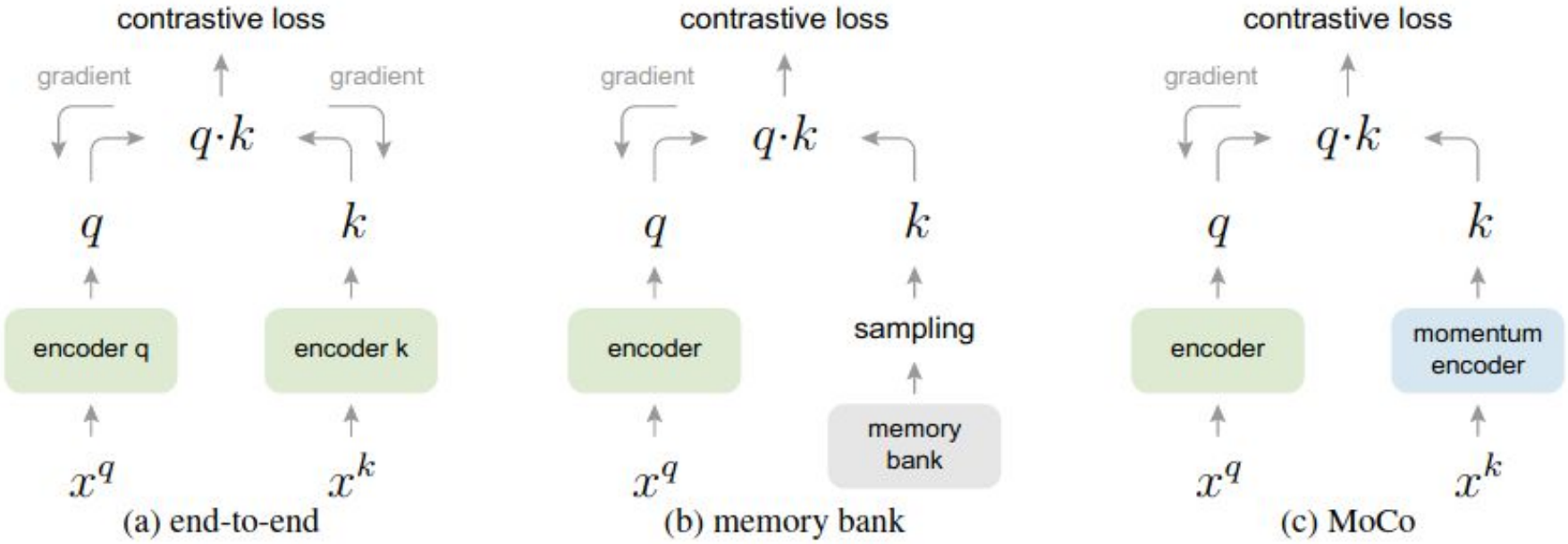
- Proxy task - Instance discrimination
- **q** - query
- **k+** - Augmented from query original image
- **k-** - Unmatching image to the query

$$\mathcal{L}_{q,k^+,\{k^-\}} = -\log \frac{\exp(q \cdot k^+/\tau)}{\exp(q \cdot k^+/\tau) + \sum_{k^-} \exp(q \cdot k^-/\tau)}$$





Conceptual comparison of three mechanisms



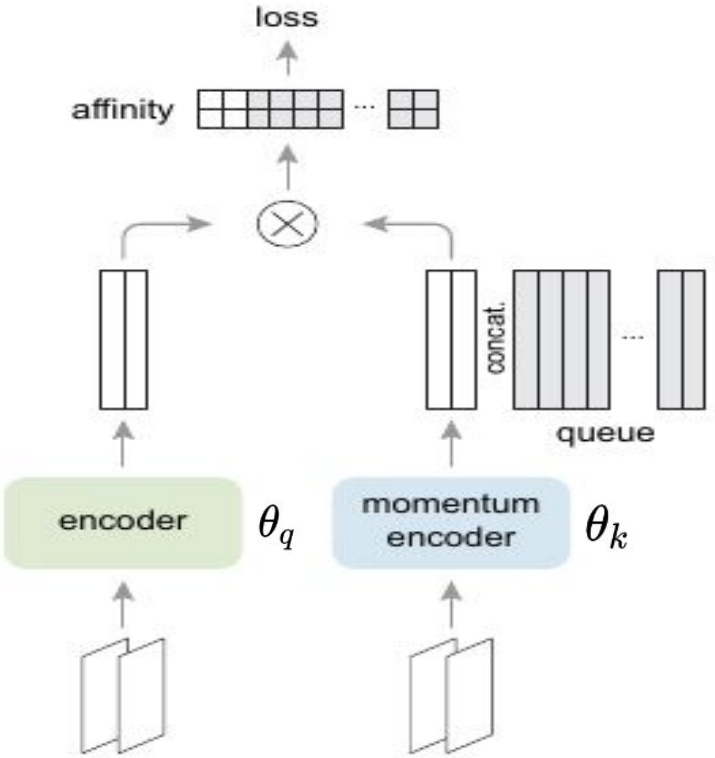
- Limited k-dim
- Inconsistent encoding
- More consistent feature encoding
- Large Memory

MoCo solution

- Encodes the keys on-the-fly
- Maintains the queue of keys
- Key encoder update:

$$\theta_k := m \cdot \theta_k + (1 - m) \cdot \theta_q$$

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



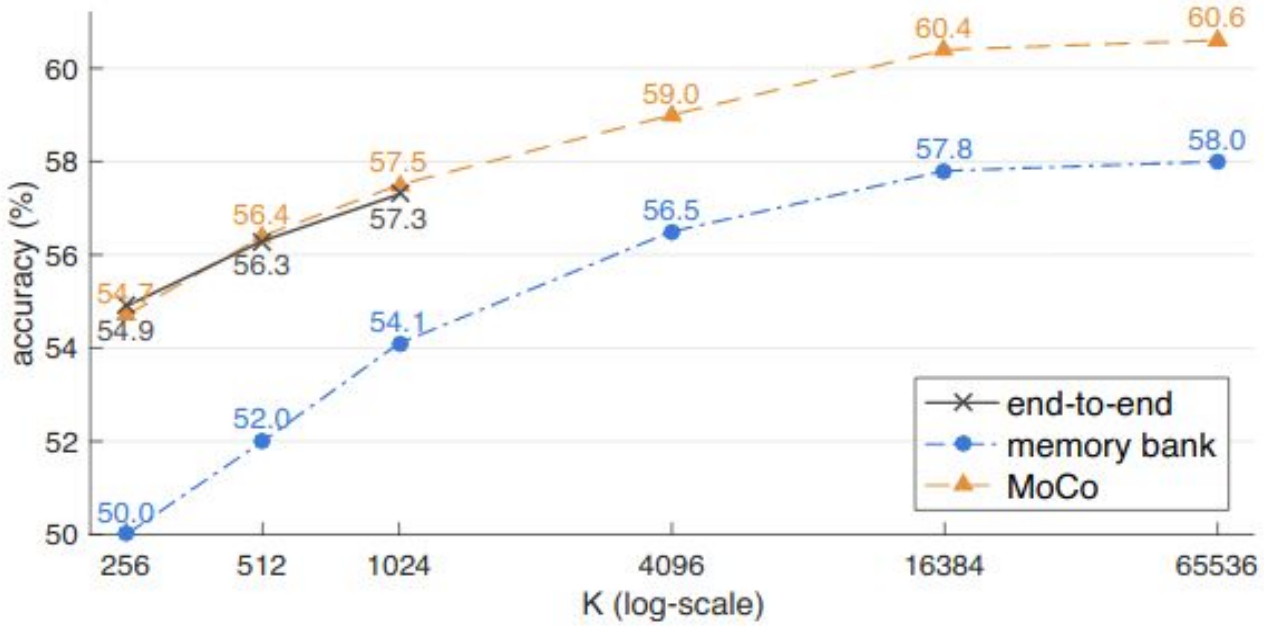
Comparison on ImageNet

- Pretext task: Instance Discrimination
- Computation: 8 x 32GB GPU

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

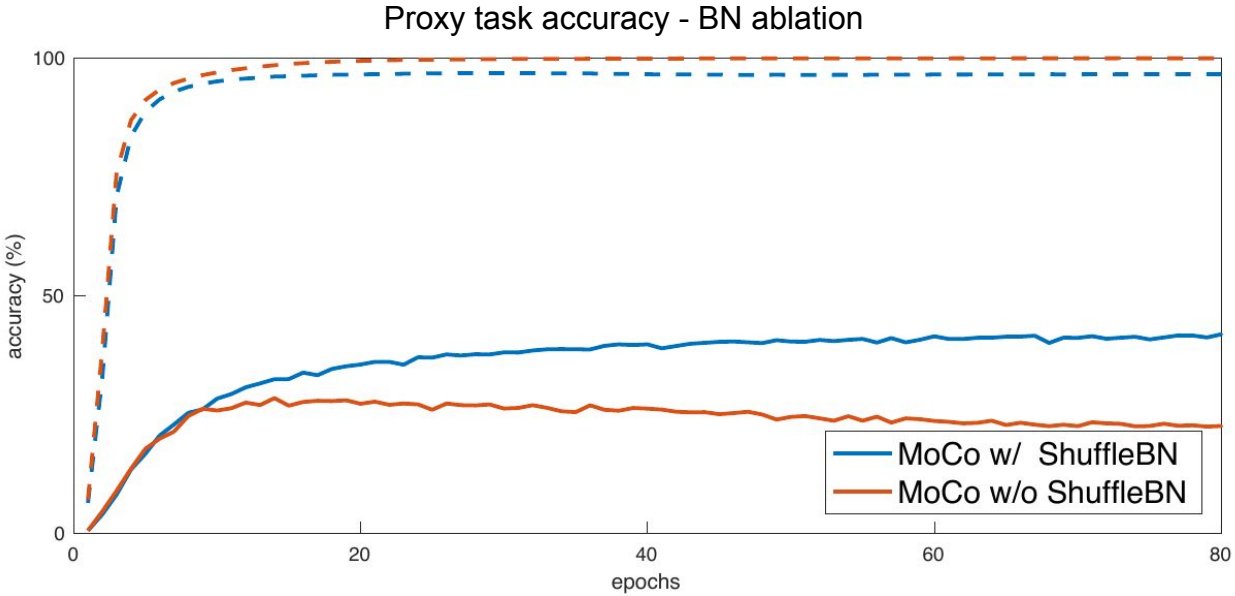
Table 3. **Memory and time cost** in 8 V100 16G GPUs

ImageNet classification



Shuffling Batch Normalization

- BN leaks intra-batch information, where positive key is
- Solution: Shuffle batch for key encoder forward pass

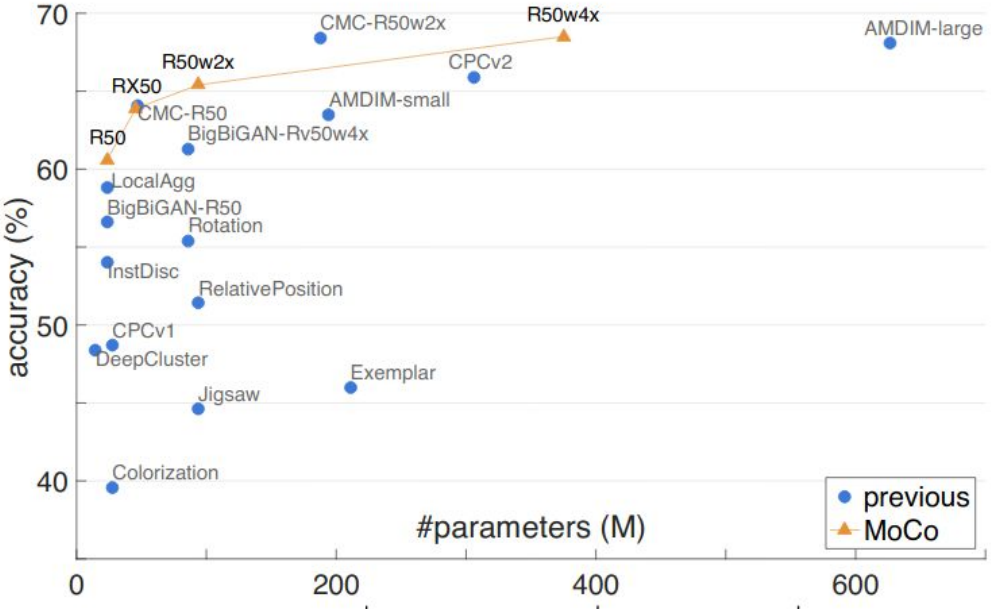


- *Dash*: Training curve
- *Solid*: Validation curve

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.

MoCo Results

Self-supervised methods on ImageNet



- IN-1M
  - ImageNet pretraining
- IG-1B
  - Instagram: 1 billion images

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**

Different task results

- MoCo can outperform ImageNet supervised pre-training in 7 vision tasks
- MoCo in IG-1B setup is consistently better than IN-1M
  - Perform well of large-scale and uncurated dataset
  - Real-world unsupervised learning setup

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)

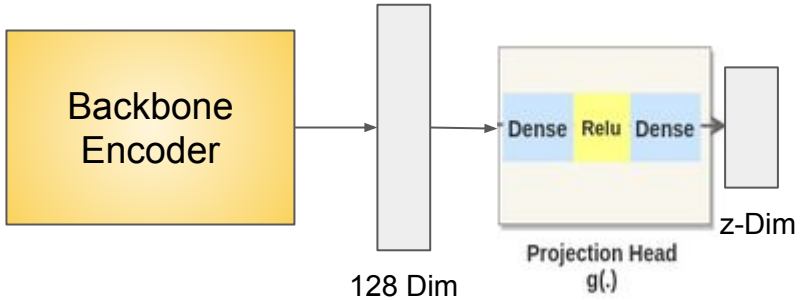
pre-train	LVIS v0.5 instance segmentation AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
random init.	22.5	34.8	23.8
super. IN-1M†	24.4	37.8	25.8
MoCo IN-1M	24.1 (−0.3)	37.4 (−0.4)	25.5 (−0.3)
MoCo IG-1B	24.9 (+0.5)	38.2 (+0.4)	26.4 (+0.6)

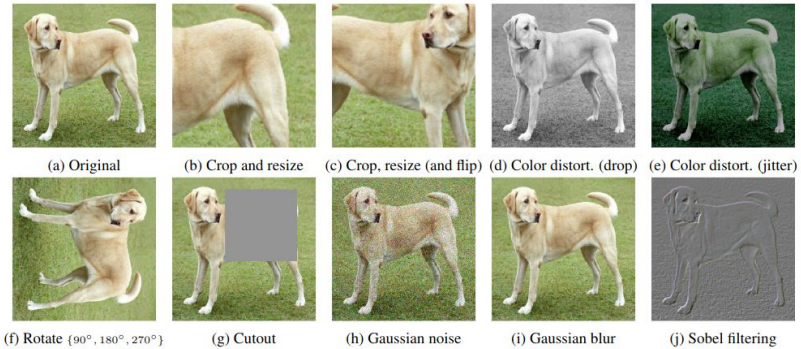
pre-train	Cityscapes instance seg. AP^{mk}	AP^{mk}_{50}	Semantic seg. (mIoU) Cityscapes	VOC
random init.	25.4	51.1	65.3	39.5
super. IN-1M	32.9	59.6	74.6	74.4
MoCo IN-1M	32.3 (−0.6)	59.3 (−0.3)	75.3 (+0.7)	72.5 (−1.9)
MoCo IG-1B	32.9 (0.0)	60.3 (+0.7)	75.5 (+0.9)	73.6 (−0.8)

Table 6: MoCo vs. ImageNet supervised pre-training, fine-tuned on various tasks

MoCo v2

- Improved Baselines with Momentum Contrastive Learning
- Combining approach from SimCLR
  - Addition of MLP (projection head)
  - heavy data augmentation



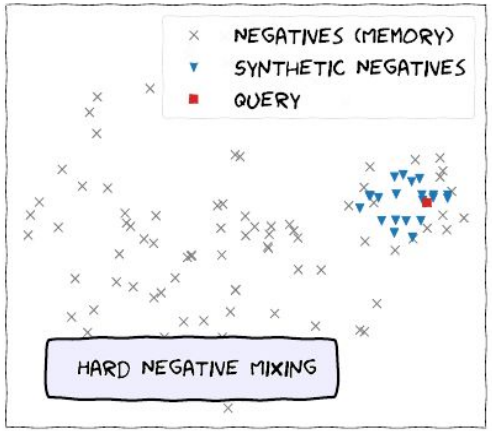


case	unsup. pre-train MLP	aug+	cos	epochs	batch	ImageNet acc.
MoCo v1 [6]				200	256	60.6
SimCLR [2]	✓	✓	✓	200	256	61.9
SimCLR [2]	✓	✓	✓	200	8192	66.6
MoCo v2	✓	✓	✓	200	256	**67.5**
*results of **longer** unsupervised training follow:*						
SimCLR [2]	✓	✓	✓	1000	4096	69.3
MoCo v2	✓	✓	✓	800	256	**71.1**

Table 2. **MoCo *vs.* SimCLR**: ImageNet linear classifier accuracy

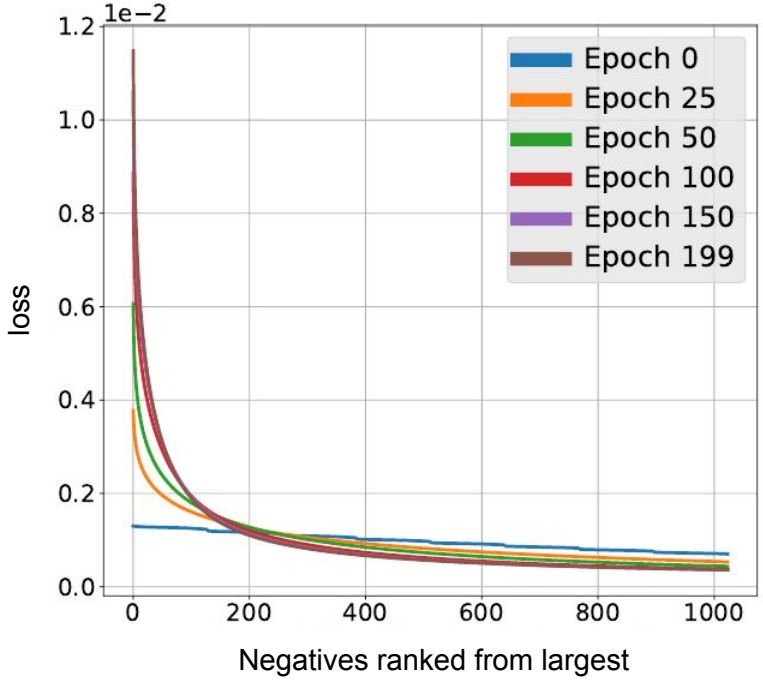
Hard Negative Mixing for Contrastive Learning

- "(M)ixing (o)f (C)ontrastive (H)ard negat(i)ves - MoCHi
- Synthesizing negative samples in representation space on-the-fly



Yannis Kalantidis, et al. NeurIPS 2020

Effect of negatives in one batch on contrastive loss



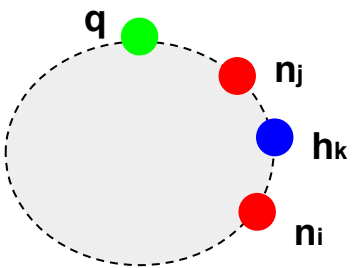
Synthesizing of Hard negatives

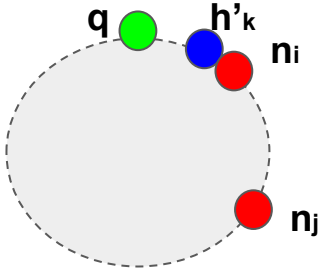
- Positive query features **q**, negative features **n**
- Convex linear combinations of pairs of its “hardest” existing negatives

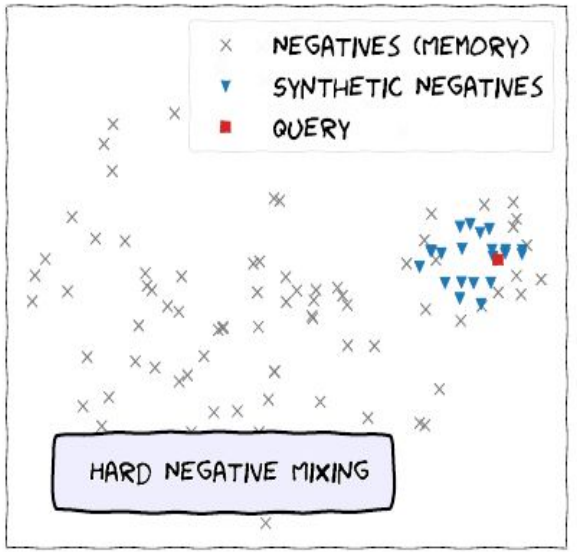
$$\mathbf{h}_k = \frac{\tilde{\mathbf{h}}_k}{\|\tilde{\mathbf{h}}_k\|_2}, \text{ where } \tilde{\mathbf{h}}_k = \alpha_k \mathbf{n}_i + (1-\alpha_k)\mathbf{n}_j,$$

- Hardest negatives from **q**

$$\mathbf{h}'_k = \frac{\tilde{\mathbf{h}}'_k}{\|\tilde{\mathbf{h}}'_k\|_2}, \text{ where } \tilde{\mathbf{h}}'_k = \beta_k \mathbf{q} + (1-\beta_k)\mathbf{n}_j$$







MoCHi Experiments

- Training a ResNet-50 model on ImageNet using 4x V100 GPU take about 6-7 days
- Consistent gains over the MoCo-v2 baseline

Synthesised from query

Hard negatives

s \ s'	0	128	256	512
0	0.0	+0.7	+0.9	+1.0
128	+0.8	+0.4	+1.1	+0.3
256	+0.3	+0.7	+0.3	+1.0
512	+0.9	+0.8	+0.6	+0.4
1024	+0.8	+1.0	+0.7	+0.6

(b) Accuracy gains over MoCo-v2 when $N = 1024$.

Method	Top1 % ($\pm\sigma$)	diff (%)
MoCo [30]	73.4	
MoCo + iMix [56]	74.2‡	↑0.8
CMC [64]	75.7	
CMC + iMix [56]	75.9‡	↑0.2
MoCo [30]*	74.0	
MoCo-v2 [13]*	78.0 ($\pm$0.2)	
+ MoCHi (1024, 1024, 128)	**79.0** ($\pm$0.4)	**↑1.0**
+ MoCHi (1024, 256, 512)	**79.0** ($\pm$0.4)	**↑1.0**
+ MoCHi (1024, 128, 256)	**78.9** ($\pm$0.5)	**↑0.9**
Using Class Oracle		
MoCo-v2*	81.8	
+ MoCHi (1024, 1024, 128)	82.5	
Supervised (Cross Entropy)	*86.2*	

Table 1: Results on ImageNet-100 after training for 200 epochs. The bottom section reports results when using a class oracle (see Section 3.3). * denotes reproduced results, ‡ denotes results visually extracted from Figure 4 in [56]. The parameters of MoCHi are (N, s, s').

Different task results

Method	IN-1k Top1	VOC 2007 AP_{50}	VOC 2007 AP	VOC 2007 AP_{75}
100 epoch training				
MoCo-v2 [13]*	63.6	80.8 (±0.2)	53.7 (±0.2)	59.1 (±0.3)
+ MoCHi (256, 512, 0)	63.9	81.1 (±0.1) (↑0.4)	54.3 (±0.3) (↑0.7)	60.2 (±0.1) (↑1.2)
+ MoCHi (256, 512, 256)	63.7	**81.3** (±0.1) (**↑0.6**)	54.6 (±0.3) (↑1.0)	60.7 (±0.8) (↑1.7)
+ MoCHi (128, 1024, 512)	63.4	81.1 (±0.1) (↑0.4)	**54.7** (±0.3) (**↑1.1**)	**60.9** (±0.1) (**↑1.9**)
200 epoch training				
MoCo-v2 [13]*	67.9	82.5 (±0.2)	56.8 (±0.1)	63.3 (±0.4)
+ MoCHi (1024, 512, 256)	68.0	82.3 (±0.2) (↓0.2)	56.7 (±0.2) (↓0.1)	63.8 (±0.2) (↑0.5)
+ MoCHi (512, 1024, 512)	67.6	82.7 (±0.1) (↑0.2)	57.1 (±0.1) (↑0.3)	64.1 (±0.3) (↑0.8)
+ MoCHi (256, 512, 0)	67.7	**82.8** (±0.2) (↑0.3)	57.3 (±0.2) (↑0.5)	64.1 (±0.1) (↑0.8)
+ MoCHi (256, 512, 256)	67.6	82.6 (±0.2) (↑0.1)	57.2 (±0.3) (↑0.4)	64.2 (±0.5) (↑0.9)
+ MoCHi (256, 2048, 2048)	*67.0*	82.5 (±0.1) (0.0)	57.1 (±0.2) (↑0.3)	64.4 (±0.2) (↑1.1)
+ MoCHi (128, 1024, 512)	*66.9*	82.7 (±0.2) (↑0.2)	57.5 (±0.3) (↑0.7)	64.4 (±0.4) (↑1.1)
Supervised [30]	76.1	81.3	53.5	58.8

MoCo and MoCHi Comparison

- MoCHi does not show performance gains over MoCo-v2 for linear classification on ImageNet-1K
- Model learn faster with MoCHi and achieves performance gains over MoCo-v2 for transfer learning
  - In 200 epochs MoCHi can achieve performance similar to MoCo-v2 after 800 epochs on PASCAL VOC
- Performance gains of MoCHi are consistent across multiple configurations
- Both methods outperforms its supervised pre-training counterpart in 7 detection/segmentation tasks

Summary and conclusion

- Identified the need for harder negatives
- Provides more generalizable feature representations
- Considerable gains without extensive hyperparameters searches
- These approaches can be implemented on top of any contrastive learning loss that involves a set of negatives
- Highly computationally demanding

- *Rethinking ImageNet pre-training: K. He, et al.*

References

Momentum Contrast for Unsupervised Visual Representation Learning:
https://arxiv.org/abs/1911.05722, CVPR 2020

Hard Negative Mixing for Contrastive Learning

https://arxiv.org/pdf/2010.01028.pdf, NeurIPS 2020

A Simple Framework for Contrastive Learning of Visual representations
https://arxiv.org/abs/2002.05709, ICML 2020

Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination

https://arxiv.org/pdf/1805.01978.pdf , CVPR 2018

Improved Baselines with Momentum Contrastive Learning

https://arxiv.org/pdf/2003.04297.pdf, Technical report