Momentum Contrast for Unsupervised Visual Representation Learning

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross Girshick
Facebook AI Research (FAIR)
Background

- Unsupervised representation learning
  - highly **successful** in natural language processing
  - generally **lag behind** in computer vision

- Approaches related to the **contrastive loss** show promising results.
  - Build dynamic **dictionaries**
  - Trains **encoders** to perform dictionary look-up
  - An encoded “**query**” (images or patches) should be **similar** to its matching key and **dissimilar** to others
Method

- Hypothesize: the dictionary should be **large** and **consistent**
- This paper presents **Momentum Contrast (MoCo)** as a way of building **large and consistent dictionaries** for unsupervised learning with a contrastive loss
Method

(K+1)-way softmax-based classifier

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

momentum update with the query encoder (m=0.99)

$$\theta_k \leftarrow m \theta_k + (1 - m) \theta_q$$

The dictionary is built as a queue, with the current mini-batch enqueued and the oldest mini-batch dequeued.
Relations to previous mechanisms

Encoder q and k are different.

The representation of a sample in memory bank is updated when it was last seen.

compare with (a) → consistent
compare with (b) → large
Experiment

- **Answer two-fold questions**
  - comparison of three mechanisms
  - performance of downstream tasks

- **Dataset**
  - **ImageNet-1M (IN-1M)**: ~1.28 million images in 1000 classes
  - **Instagram-1B (IG-1B)**: ~1 billion (940M) public images from ~1500 hashtags (long-tailed, unbalanced distribution)
Compare three mechanisms

- linear classification on frozen features
Compare with previous methods

| 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         |

Key observations:
- higher accuracy with less #parameters
- efficiency, less #parameters with higher accuracy
performance of downstream tasks

| pre-train          | RelPos, by [14] | Multi-task [14] | Jigsaw, by [26] | LocalAgg [66] | MoCo | AP | MoCo | 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) |

Table 4. Comparison with previous methods on object detection fine-tuned on PASCAL VOC trainval2007. Evaluation is on
The performance of downstream tasks is shown in the table below:

| pre-train | COCO keypoint detection | LVIS v0.5 instance segmentation |
|-----------|--------------------------|---------------------------------|
|           | AP<sub>kp</sub> | AP<sub>kp</sub><sub>50</sub> | AP<sub>kp</sub><sub>75</sub> | AP<sub>mk</sub> | AP<sub>mk</sub><sub>50</sub> | AP<sub>mk</sub><sub>75</sub> |
| random init. | 65.9 | 86.5 | 71.7 | 22.5 | 34.8 | 23.8 |
| super. IN-1M | 65.8 | 86.9 | 71.9 | 24.4 | 37.8 | 25.8 |
| MoCo IN-1M | 66.8 (+1.0) | 87.4 (+0.5) | 72.5 (+0.6) | 24.1 (−0.3) | 37.4 (−0.4) | 25.5 (−0.3) |
| MoCo IG-1B | 66.9 (+1.1) | 87.8 (+0.9) | 73.0 (+1.1) | 24.9 (+0.5) | 38.2 (+0.4) | 26.4 (+0.6) |
|           | COCO dense pose estimation | Cityscapes instance seg. | Semantic seg. (mIoU) |
|           | AP<sub>dp</sub> | AP<sub>dp</sub><sub>50</sub> | AP<sub>dp</sub><sub>75</sub> | AP<sub>mk</sub> | AP<sub>mk</sub><sub>50</sub> | Ap<sub>mk</sub><sub>75</sub> | Cityscapes | VOC |
| random init. | 39.4 | 78.5 | 35.1 | 25.4 | 51.1 | 65.3 | 39.5 |
| super. IN-1M | 48.3 | 85.6 | 50.6 | 32.9 | 59.6 | 74.6 | 74.4 |
| MoCo IN-1M | 50.1 (+1.8) | 86.8 (+1.2) | 53.9 (+3.3) | 32.3 (−0.6) | 59.3 (−0.3) | 75.3 (+0.7) | 72.5 (−1.9) |
| MoCo IG-1B | 50.6 (+2.3) | 87.0 (+1.4) | 54.3 (+3.7) | 32.9 (0.0) | 60.3 (+0.7) | 75.5 (+0.9) | 73.6 (−0.8) |
Further reading

- A Simple Framework for Contrastive Learning of Visual Representations
  - Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton (Google Brain)

- Learning deep representations by mutual information estimation and maximization
  - R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio (ICLR 2019)

- Unsupervised feature learning via non-parametric instance discrimination
  - Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin (CVPR 2018 spotlight)