# Momentum Contrast for Unsupervised Visual Representation Learning


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


## CVPR 2020

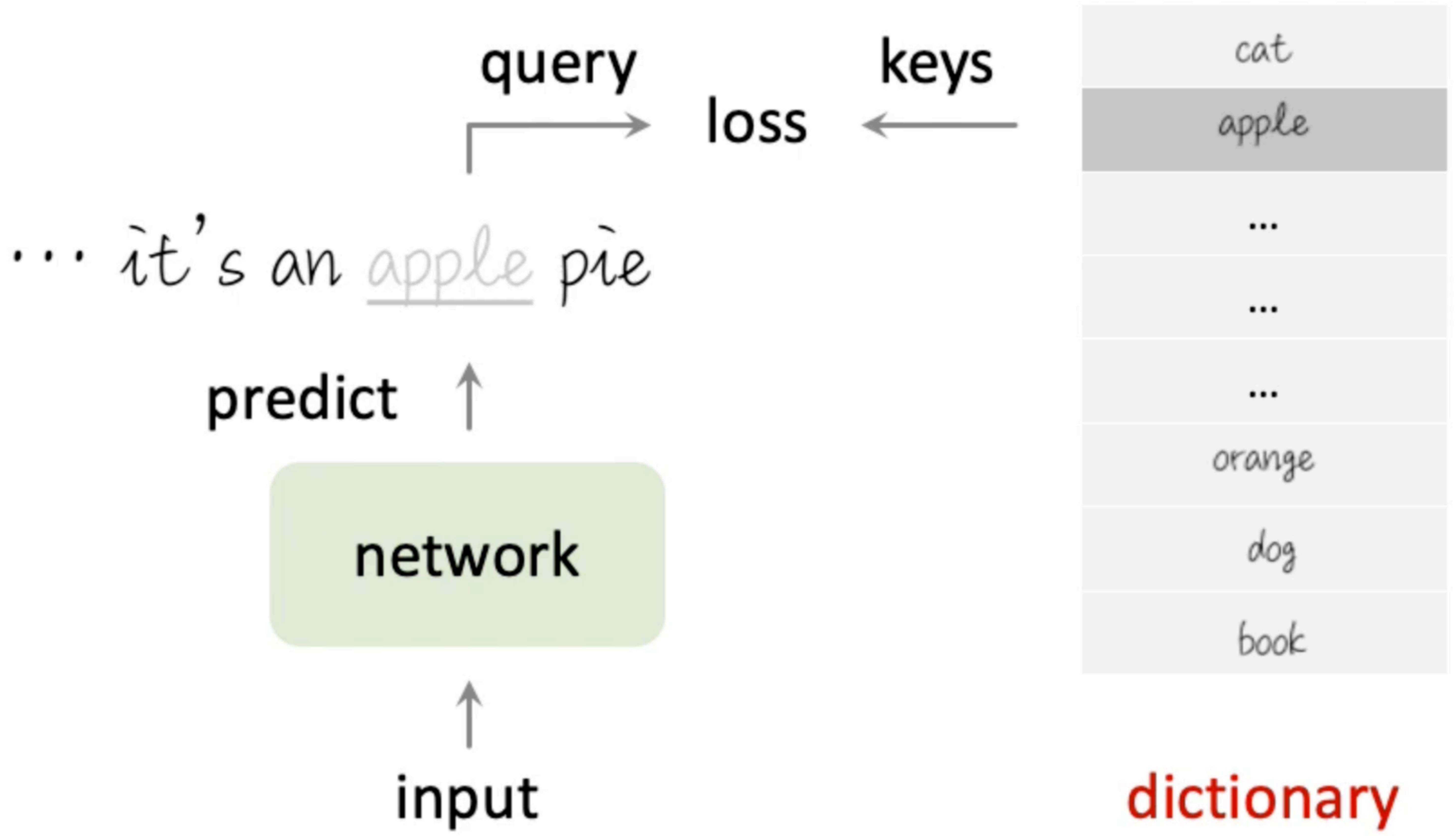
**Cheng Jiang, EECS 598-012 Unsupervised Visual Learning**
**February 8, 2021**

# Background

Unsupervised representation learning is successful in natural language processing (NLP), but lags behind for visual learning

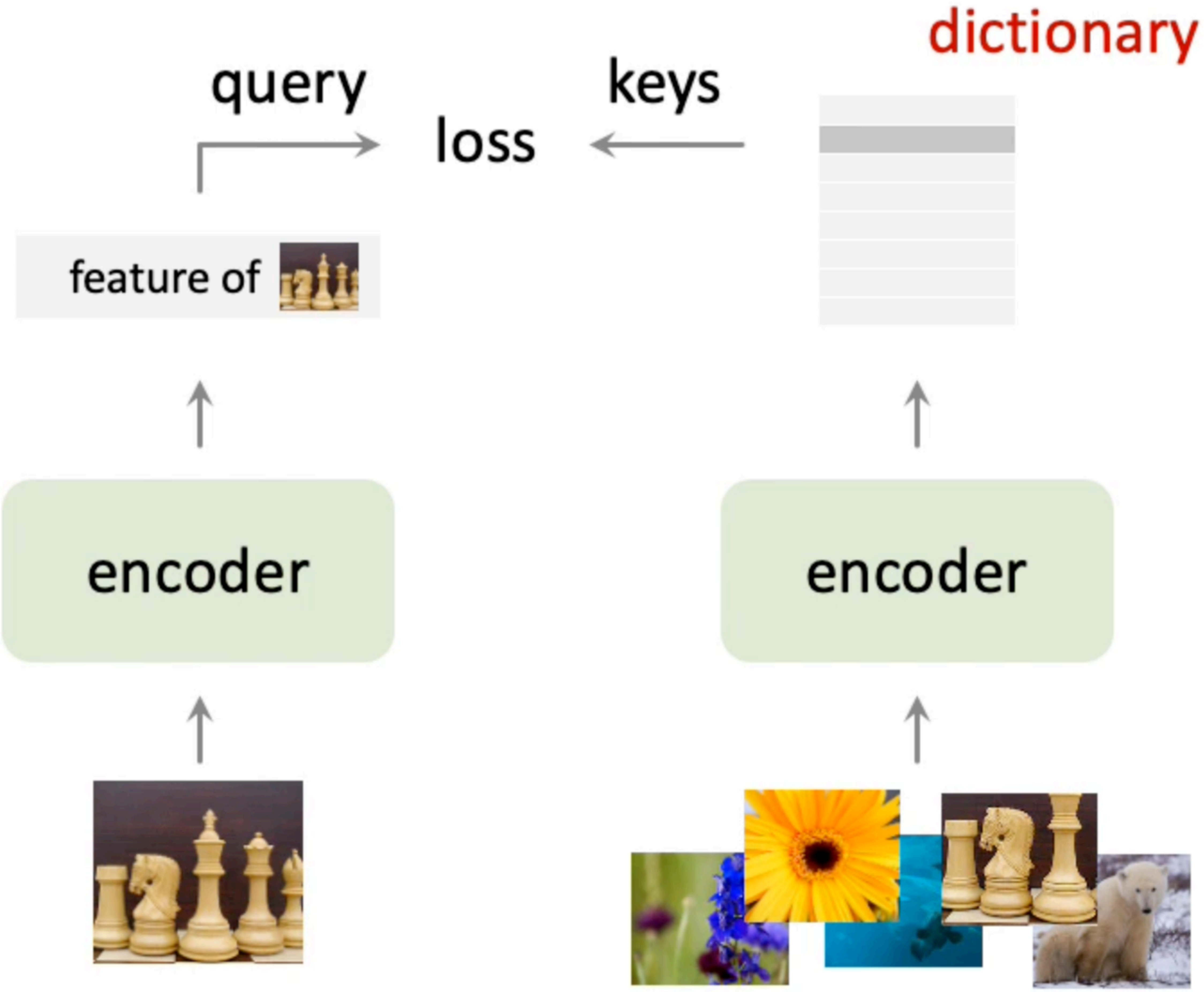
| | Tokenized Dictionaries | Signal Space |
|---|---|---|
| Language Tasks | Word → Representation | Discrete |
| Visual Learning | Image Samples → Representation | Continuous, High-dimensional |

# Contrastive Learning in NLP




Devlin et al. NAACL 2019 (BERT)
Image credit: He et al. CVPR 2020 (MoCo)

# Contrastive Learning as Dictionary Look-up




He et al. CVPR 2020 (MoCo)

**Hypothesis:** we want dictionaries that are **large** and **consistent**

# Notation: Queries, Keys, and Encoders

$$\underbrace{q}_{\text{Encoded Query}} = \underbrace{f_{\mathrm{q}}}_{\text{Query Encoder}}(\underbrace{x^q}_{\text{Input Query}})$$

$$\underbrace{k}_{\text{Encoded Key}} = \underbrace{f_{\mathrm{k}}}_{\text{Key Encoder}}(\underbrace{x^k}_{\text{Key in Dictionary}})$$

# InfoNCE Loss

$$\mathcal{L}_q = -\log \frac{\exp(\underbrace{q}_{\text{Encoded Query}} \cdot \underbrace{k_+}_{\text{Encoded Positive Key}} / \underbrace{\tau}_{\text{Temperature}})}{\sum_{i=0}^{K} \exp(\underbrace{q}_{\text{Encoded Query}} \cdot \underbrace{k_i}_{\text{(All) Encoded Keys}} / \underbrace{\tau}_{\text{Temperature}})}$$

# Momentum Contrast Key Design Choices

- Maintain the dictionary as a queue of data samples
- Query encoder updated using momentum moving averages

Momentum

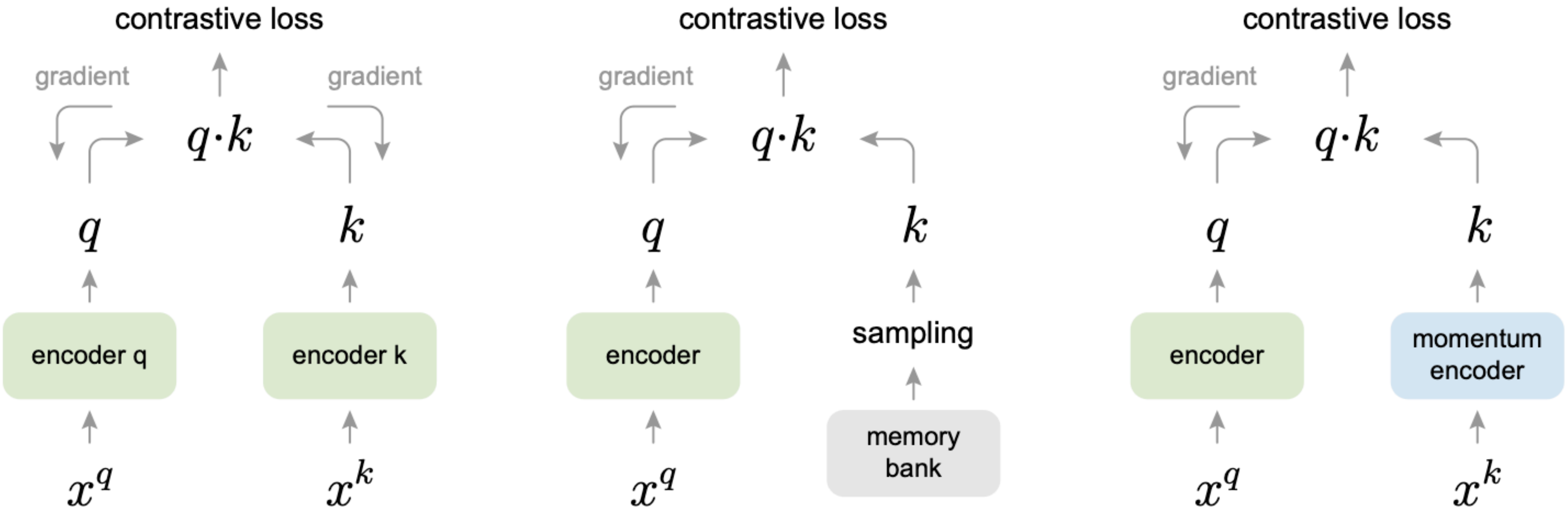
$$\theta_{\mathrm{k}} \leftarrow m\theta_{\mathrm{k}} + (1 - m)\theta_{\mathrm{q}}$$

Query Encoder Parameters

Query Encoder Parameters

- Use instance discrimination (different views of the same image) as pretext task

# Compare to Existing Mechanisms




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



He et al. CVPR 2020 (MoCo)

# Experimentation & Evaluation

Evaluation Strategies:

- Linear classification: linear classifier trained with frozen pre-trained weights

- Features fine-tuning: all layers are fine-tuned end to end

Dataset used:

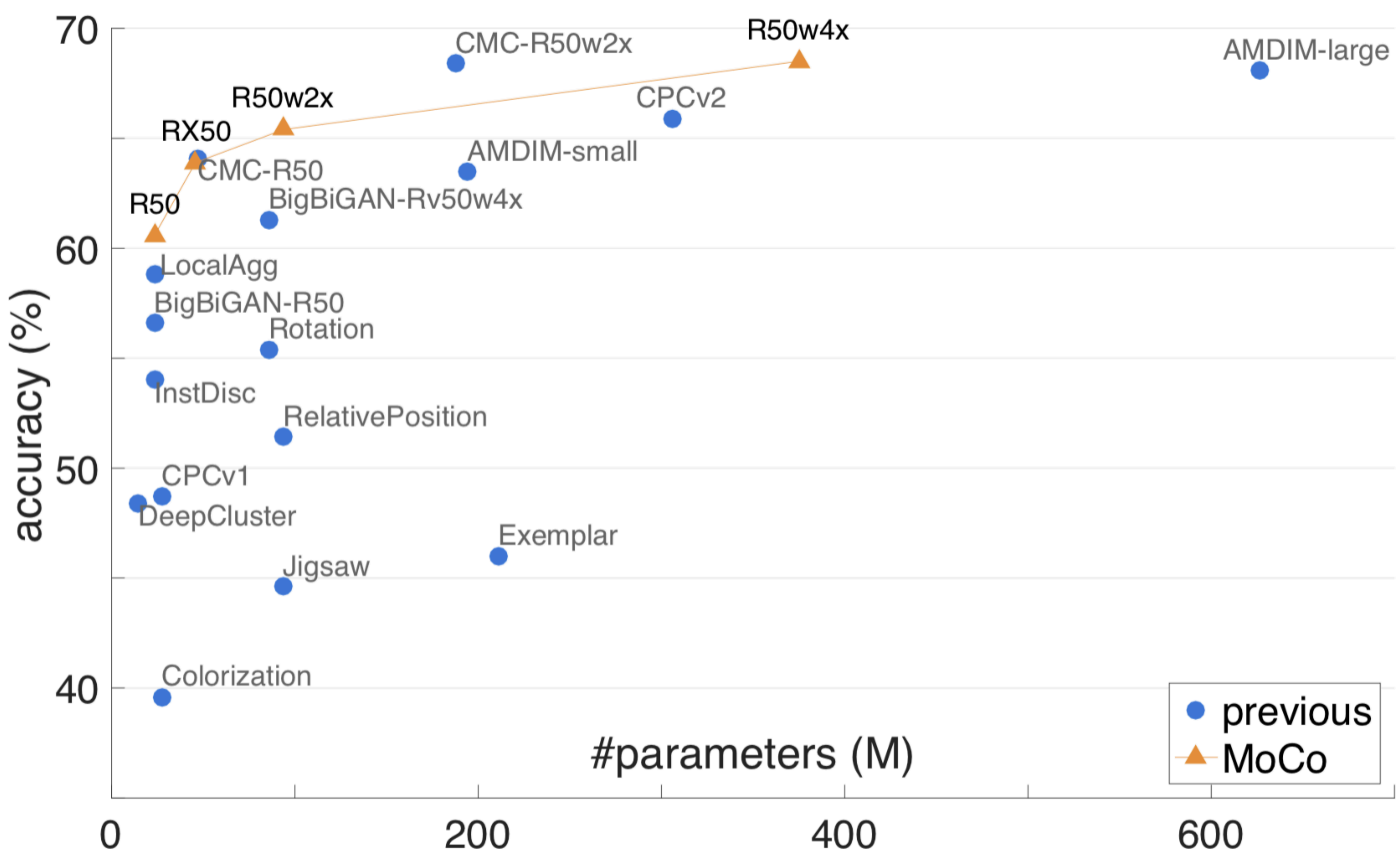
- ImageNet
  1.25M in 1000 classes
  Well balanced

- Instagram
  940M in 1500 hashtags
  Uncurated and unbalanced

# Experiment Details

- ResNet as backbone, SGD optimizer
- Feature normalized using L2-Norm
- Random crop, color jittering, horizontal flip, grayscale conversion
- Shuffling batch normalization

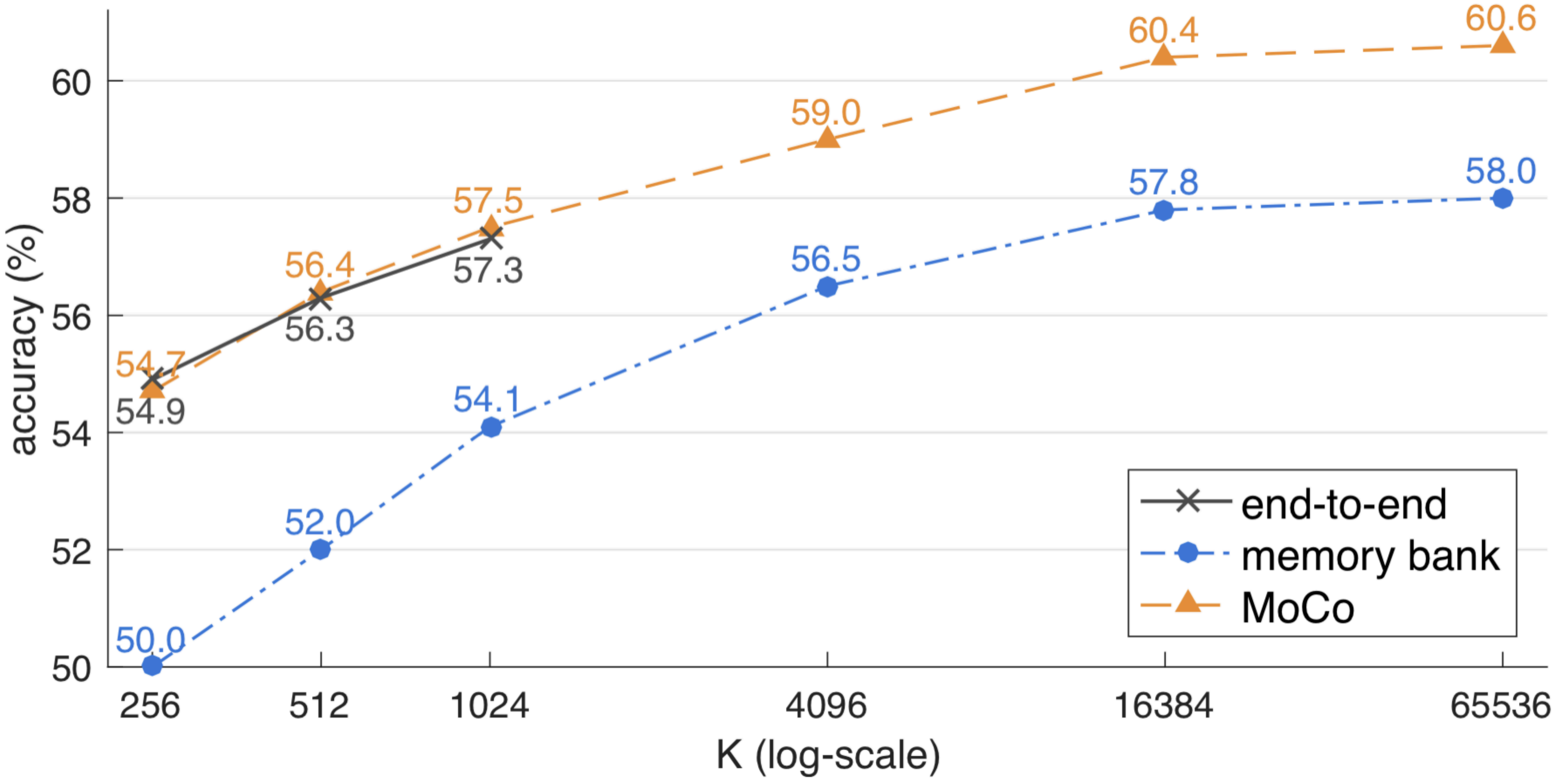
# Linear Classification Results






He et al. CVPR 2020 (MoCo)

# Ablation: Contrastive Loss Mechanisms






He et al. CVPR 2020 (MoCo)

# Ablation: Momentum

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

# Transferring Features: PASCAL VOC Object Detection

| pre-train | AP$_{50}$ | | | | | AP | AP$_{75}$ | |
|---|---|---|---|---|---|---|---|---|
| | RelPos, by [14] | Multi-task [14] | Jigsaw, by [26] | LocalAgg [66] | **MoCo** | **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) |

# Ablation: Backbones

| 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**

# Ablation: Contrastive Loss Mechanisms

| pre-train | R50-dilated-C5 | | | R50-C4 | | |
|---|---|---|---|---|---|---|
| | $AP_{50}$ | 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** |

# Transferring Features: COCO Object Detection & Segmentation

| 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**) |

(c) Mask R-CNN, R50-**C4**, **1**× schedule

| 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**) |

(d) Mask R-CNN, R50-**C4**, **2**× schedule

# Discussion

Pros:

- Proposed model is simple and intuitive
- Good experimental results, closing the gap between unsupervised learning and supervised learning
- Reduced memory usage compare to existing methods

Cons:

- Still rely on a set of hand-crafted transformations

## Next Up

Bootstrap your own latent: A new approach to self-supervised Learning

# References


- He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020 (pp. 9729-9738).
- Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.