Momentum Contrast for Unsupervised Visual Representation Learning

Kaiming He Haoqi Fan Yuxin Wu Saining Xie Ross Girshick

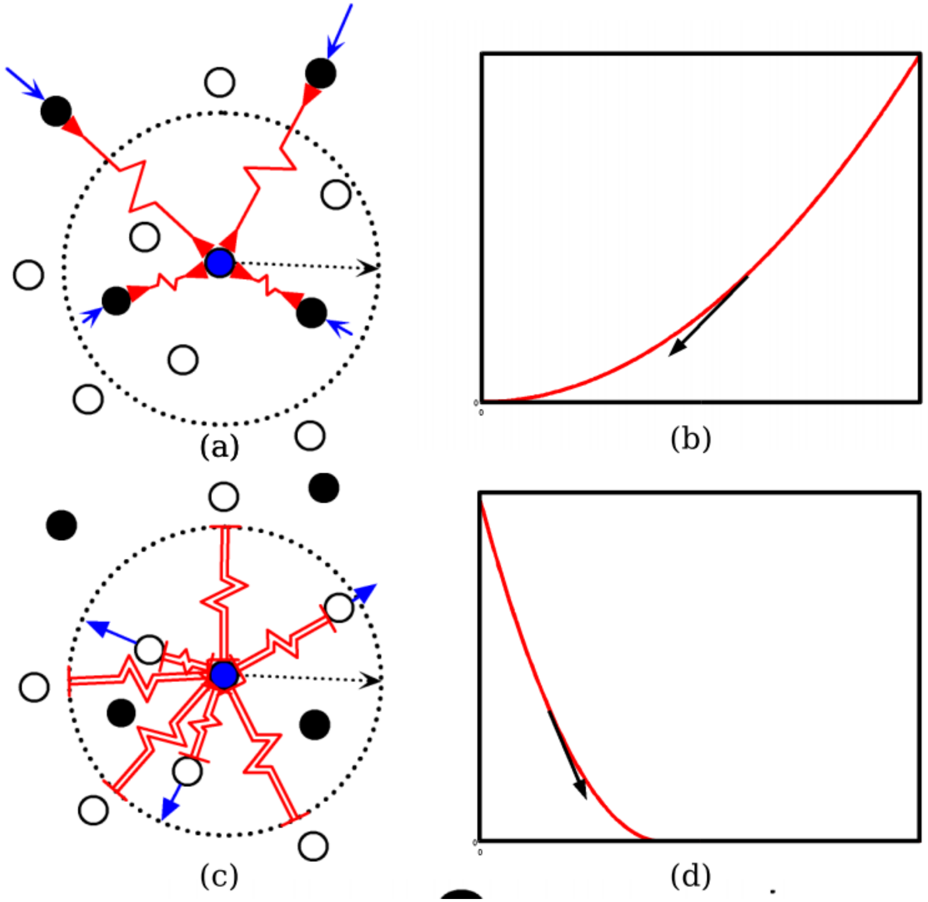
Facebook AI Research (FAIR)

Hanqing Chao

Contrastive Learning

- $Y = 0$ if X_1 and X_2 are deemed similar
- $Y = 1$ if they are deemed dissimilar

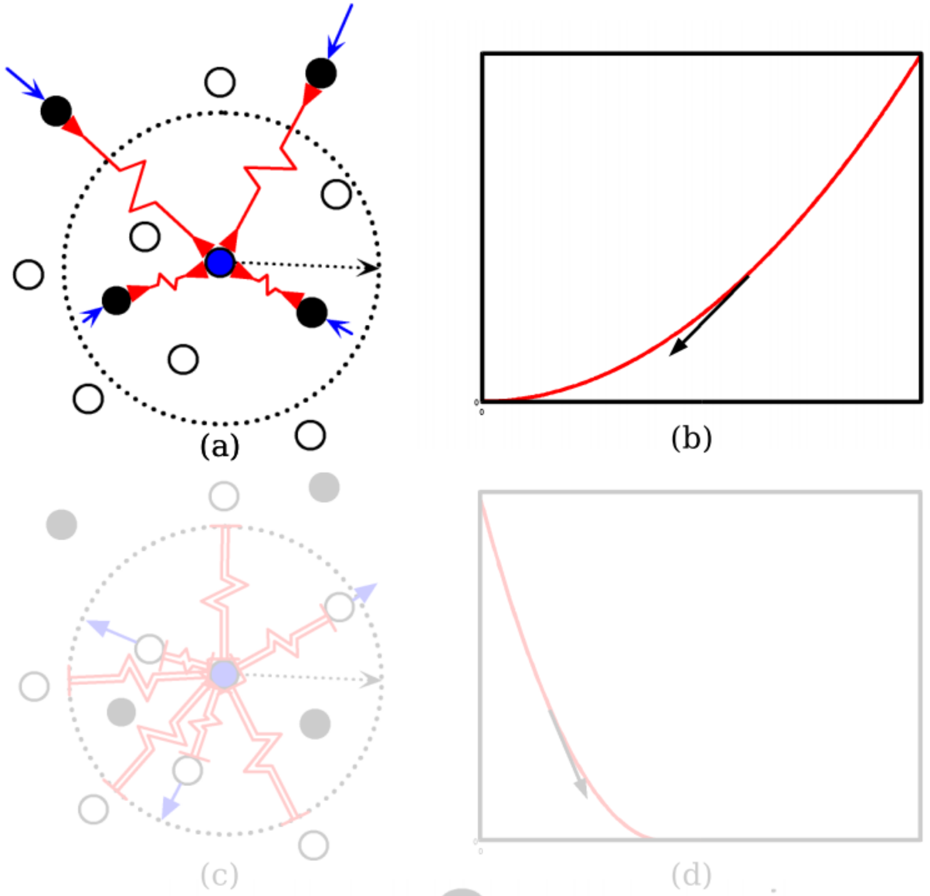
$$L(W, Y, \vec{X}_1, \vec{X}_2) =$$
$$(1 - Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{max(0, m - D_W)\}^2$$



Contrastive Learning

- $Y = 0$ if X_1 and X_2 are deemed similar
- $Y = 1$ if they are deemed dissimilar

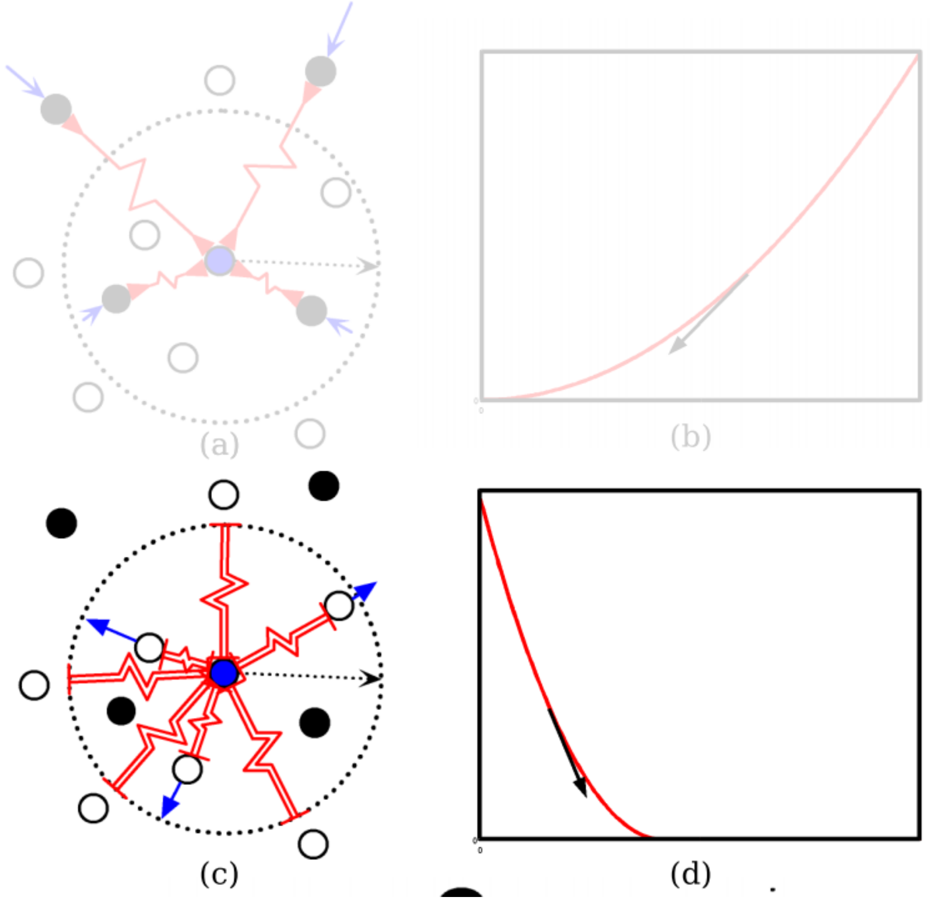
$$L(W, Y, \vec{X_1}, \vec{X_2}) =$$
$$(1-Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{max(0, m - D_W)\}^2$$



Contrastive Learning

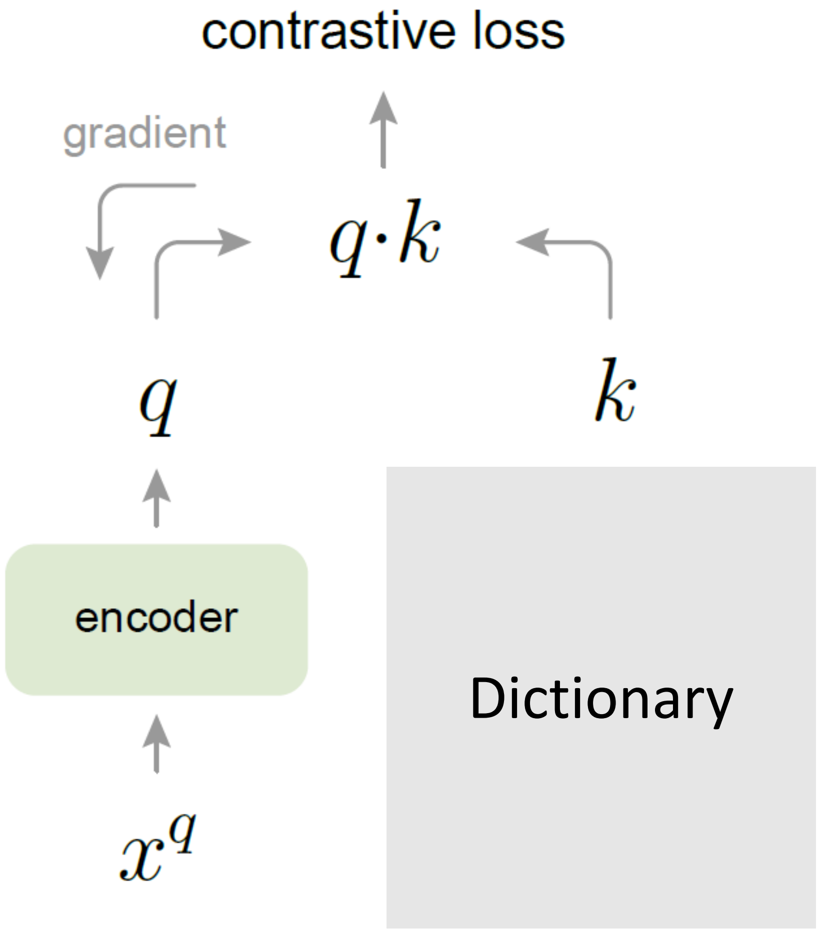
- $Y = 0$ if X_1 and X_2 are deemed similar
- $Y = 1$ if they are deemed dissimilar

$$L(W, Y, \vec{X}_1, \vec{X}_2) =$$
$$(1-Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{max(0, m - D_W)\}^2$$



Dictionary Learning

- Goal: train encoder
- To make it work, we have to find a way to establish the dictionary
- Two key properties:
  - CONSISTENCY
  - LARGE: covers a rich set of samples



Pretext task

Unsupervised learning:

With a dataset like ImageNet,

Regarding each picture as a class

Positive: if a pair of sample is generated from a same picture
Negative: otherwise

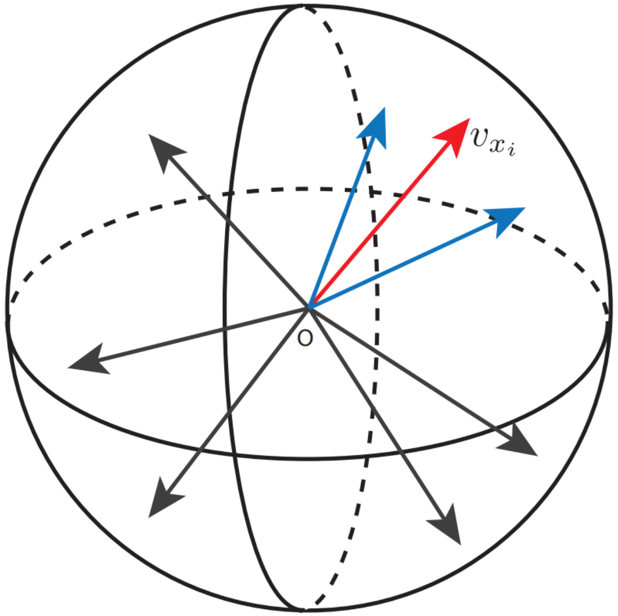
NCE

$$L(W, Y, \vec{X_1}, \vec{X_2}) = (1-Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{max(0, m - D_W)\}^2$$

Softmax:

Too much parameters

$$-\frac{1}{M}\sum_{i=1}^{M}\log\frac{e^{\boxed{W_{y_i}^T} f(\mathbf{x}_i)+b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^T f(\mathbf{x}_i)+b_j}}$$



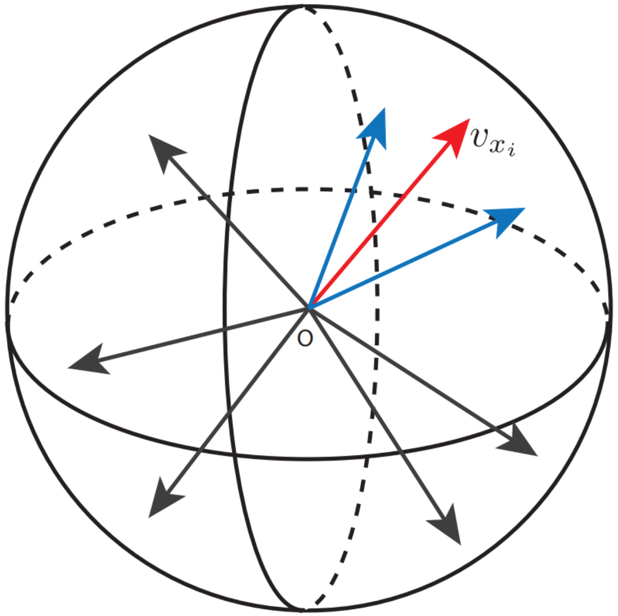
NCE

$$L(W, Y, \vec{X}_1, \vec{X}_2) = (1-Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{max(0, m - D_W)\}^2$$

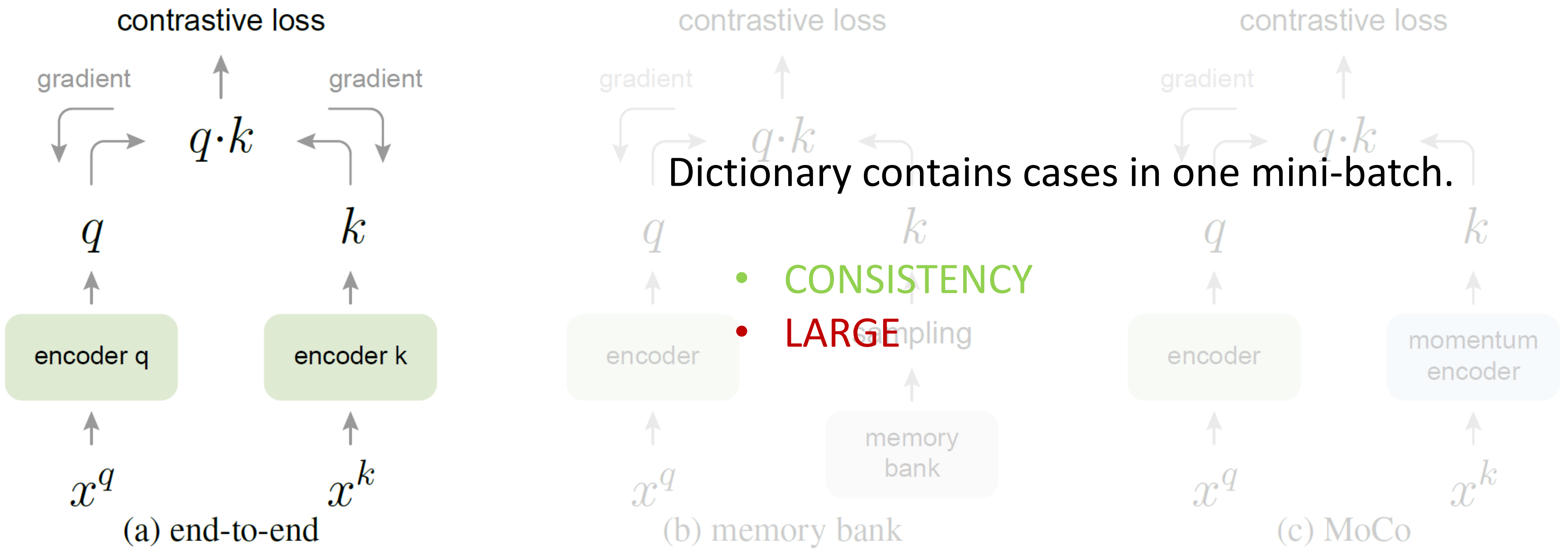
Noise-contrastive estimation:

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

Using K negative samples with 1 positive sample



Dictionary Learning



Dictionary contains cases in one mini-batch.

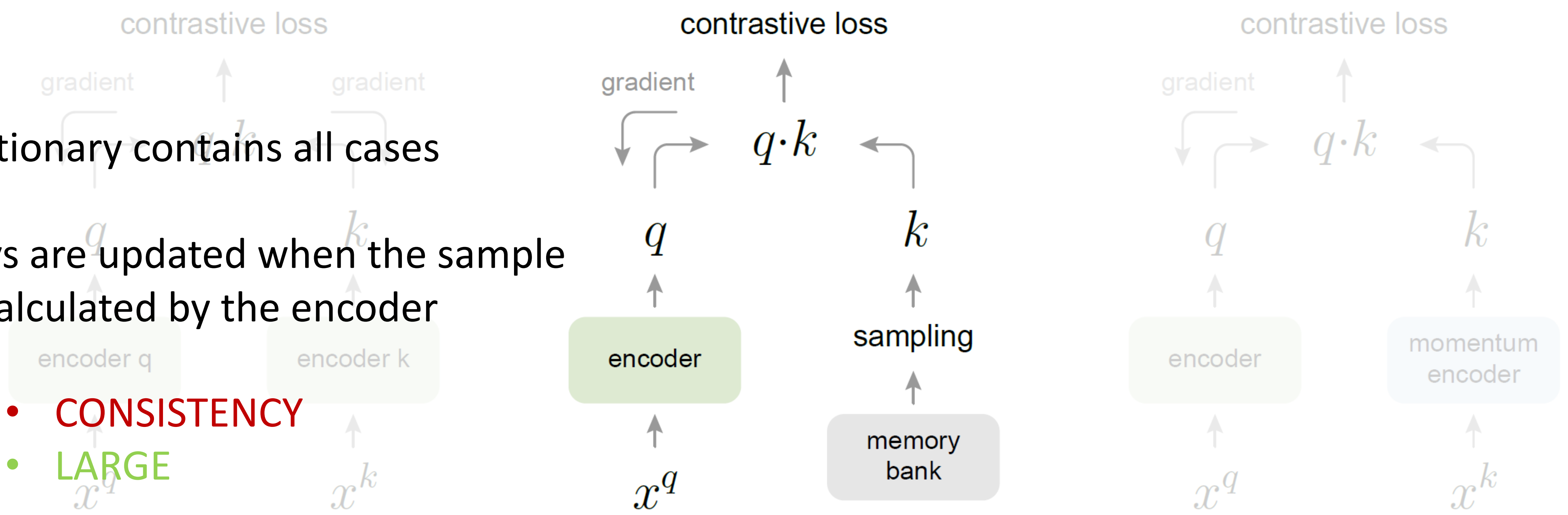
- CONSISTENCY
- LARGE

Dictionary Learning

Dictionary contains all cases

Keys are updated when the sample is calculated by the encoder

- CONSISTENCY
- LARGE

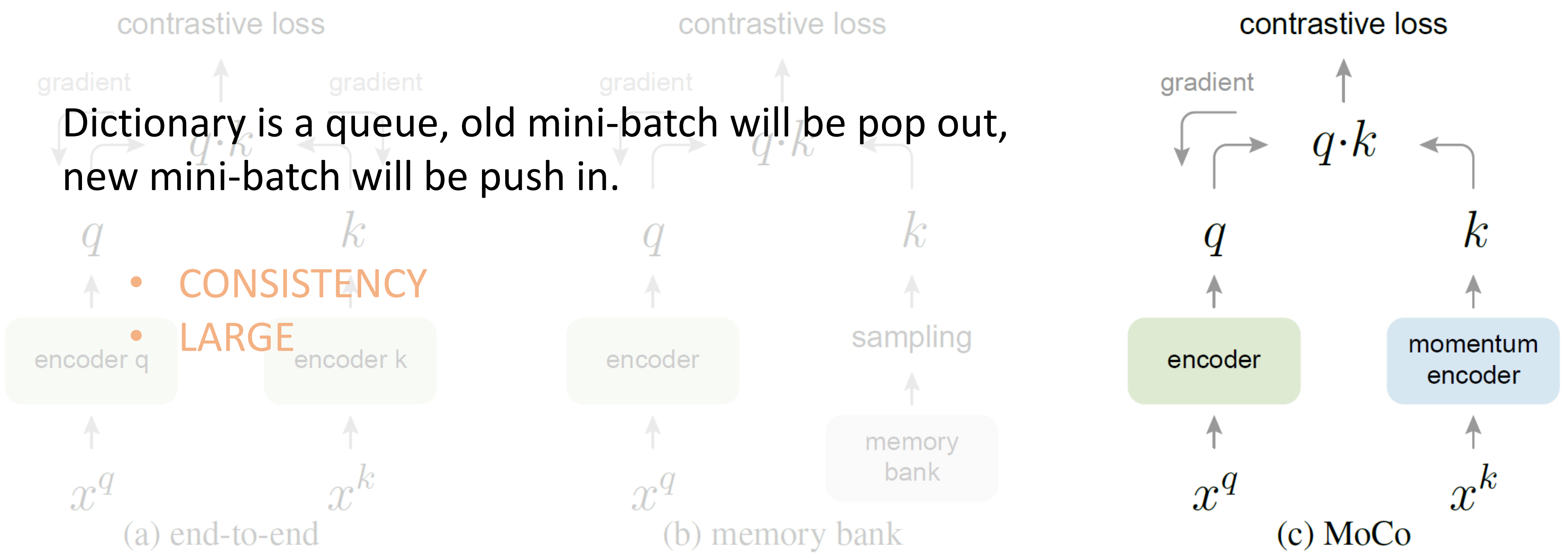


(a) end-to-end (b) memory bank (c) MoCo

Dictionary Learning

Dictionary is a queue, old mini-batch will be pop out, new mini-batch will be push in.

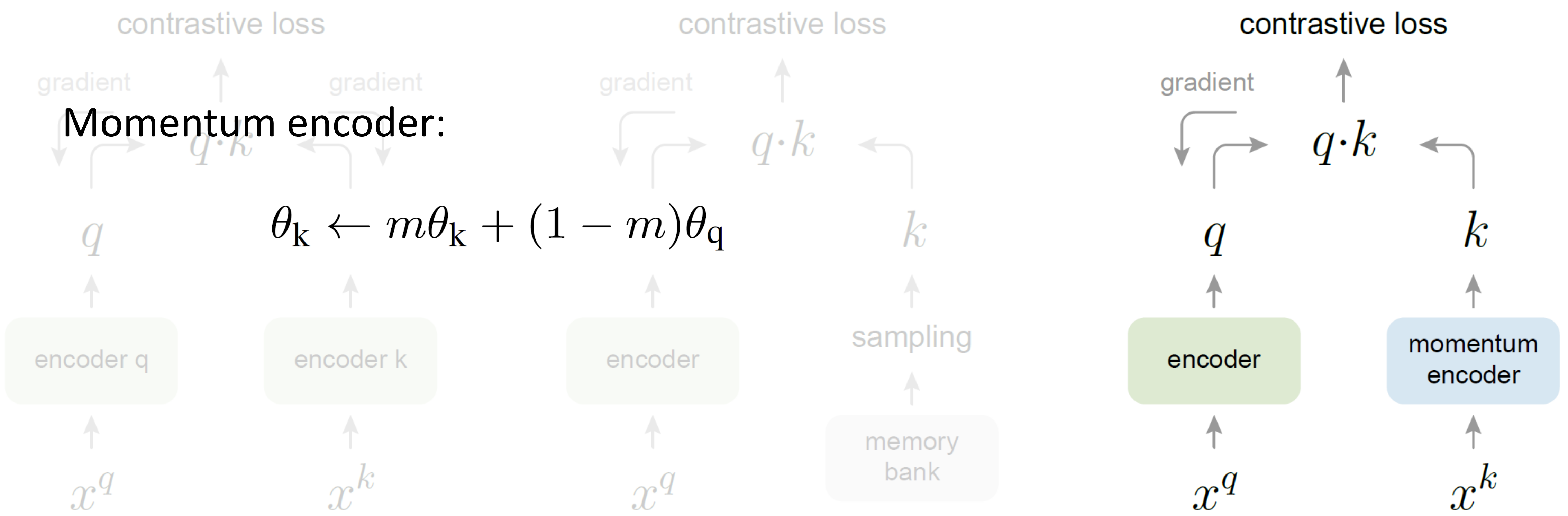
- CONSISTENCY
- LARGE



Dictionary Learning

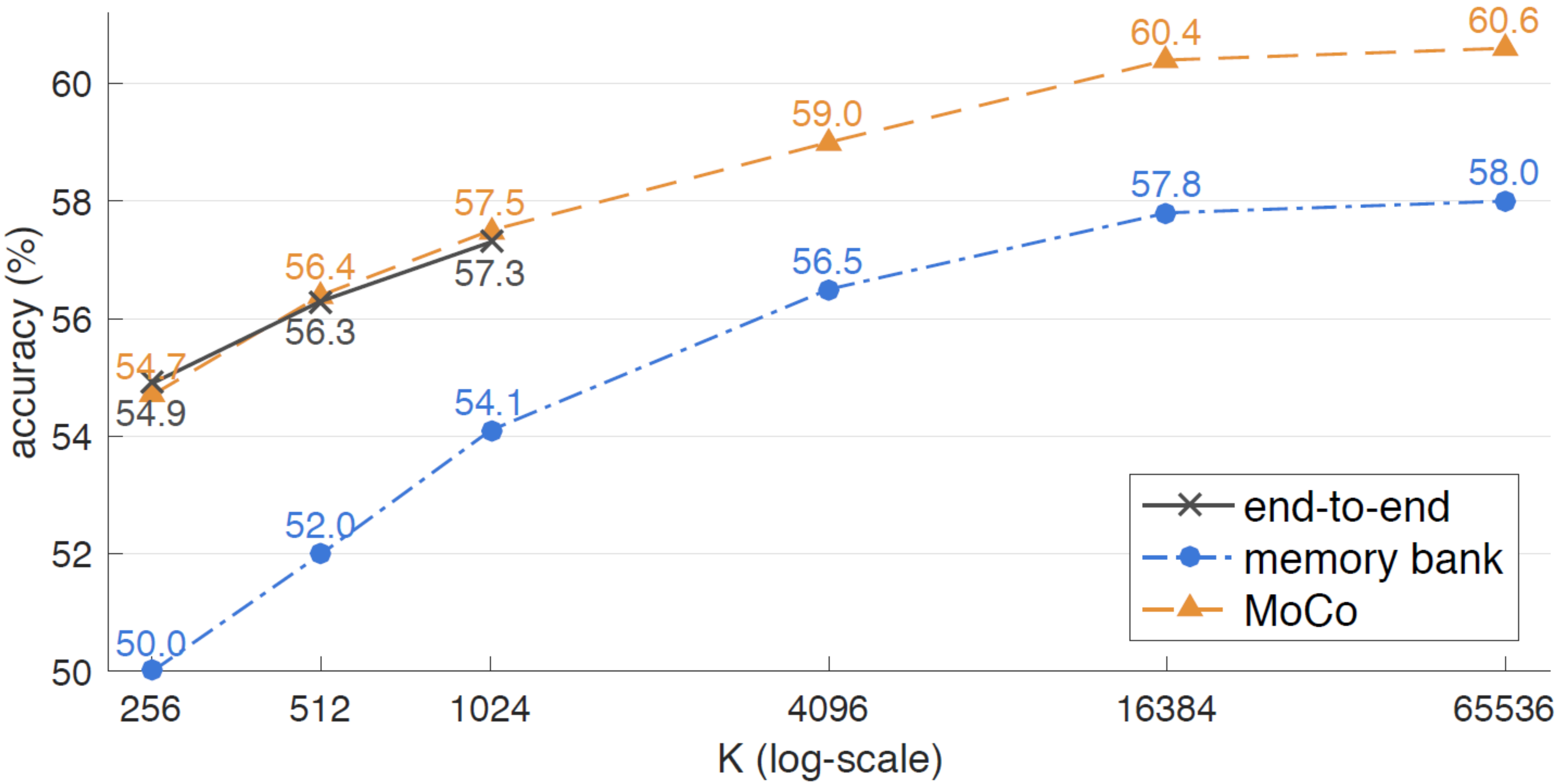
Momentum encoder:

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

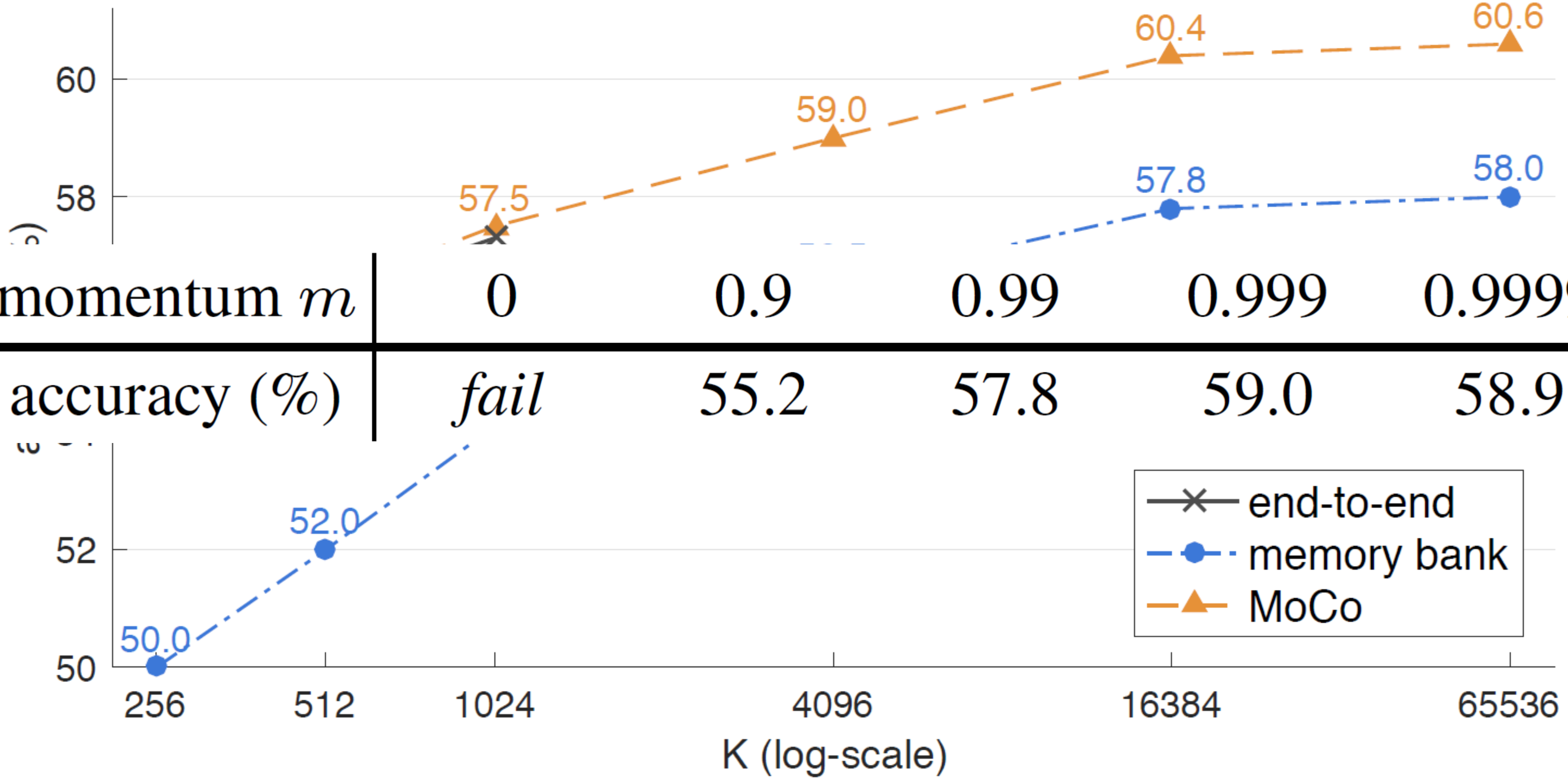


(a) end-to-end (b) memory bank (c) MoCo

ImageNet


60.6
60.4
59.0
58.0
57.8
57.5
57.3
56.5
56.4
56.3
54.7
54.9
54.1
52.0
50.0
accuracy (%)
60
58
56
54
52
50
256
512
1024
4096
16384
65536
K (log-scale)
end-to-end
memory bank
MoCo

Momentum



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

Object Detection

pre-train	AP_{50}	AP	AP_{75}
random init.	58.0	32.8	32.5
super. IN-1M	81.5	53.6	58.9
MoCo IN-1M	81.1 (−0.4)	53.8 (+0.2)	58.6 (−0.3)
MoCo IG-1B	81.6 (+0.1)	54.8 (**+1.2**)	60.3 (**+1.4**)

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

pre-train	AP_{50}	AP	AP_{75}
random init.	52.5	28.1	26.2
super. IN-1M	80.8	52.0	56.5
MoCo IN-1M	81.4 (**+0.6**)	55.2 (**+3.2**)	61.2 (**+4.7**)
MoCo IG-1B	82.1 (**+1.3**)	56.2 (**+4.2**)	62.3 (**+5.8**)

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

Object Detection

pre-train	AP_{50} RelPos, by [12]	Multi-task [12]	Jigsaw, by [24]	LocalAgg [64]	**MoCo**	AP **MoCo**	AP_{75} Multi-task [12]	**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**)

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)

(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

Thanks！