Momentum Contrast for Unsupervised Visual Representation Learning


Kaiming He


Haoqi Fan


Yuxin Wu


Saining Xie

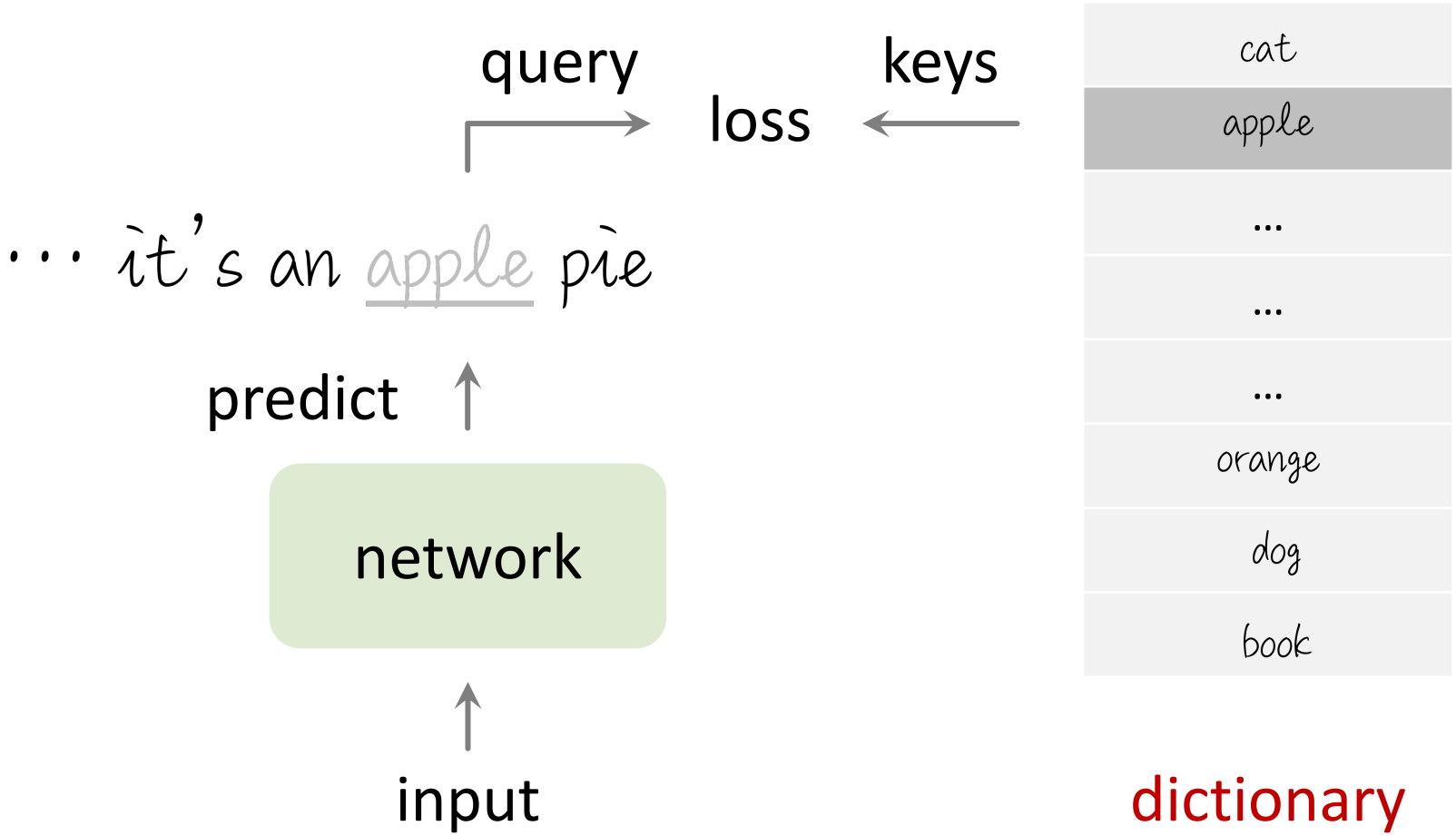

Ross Girshick

Facebook AI Research

Highlights

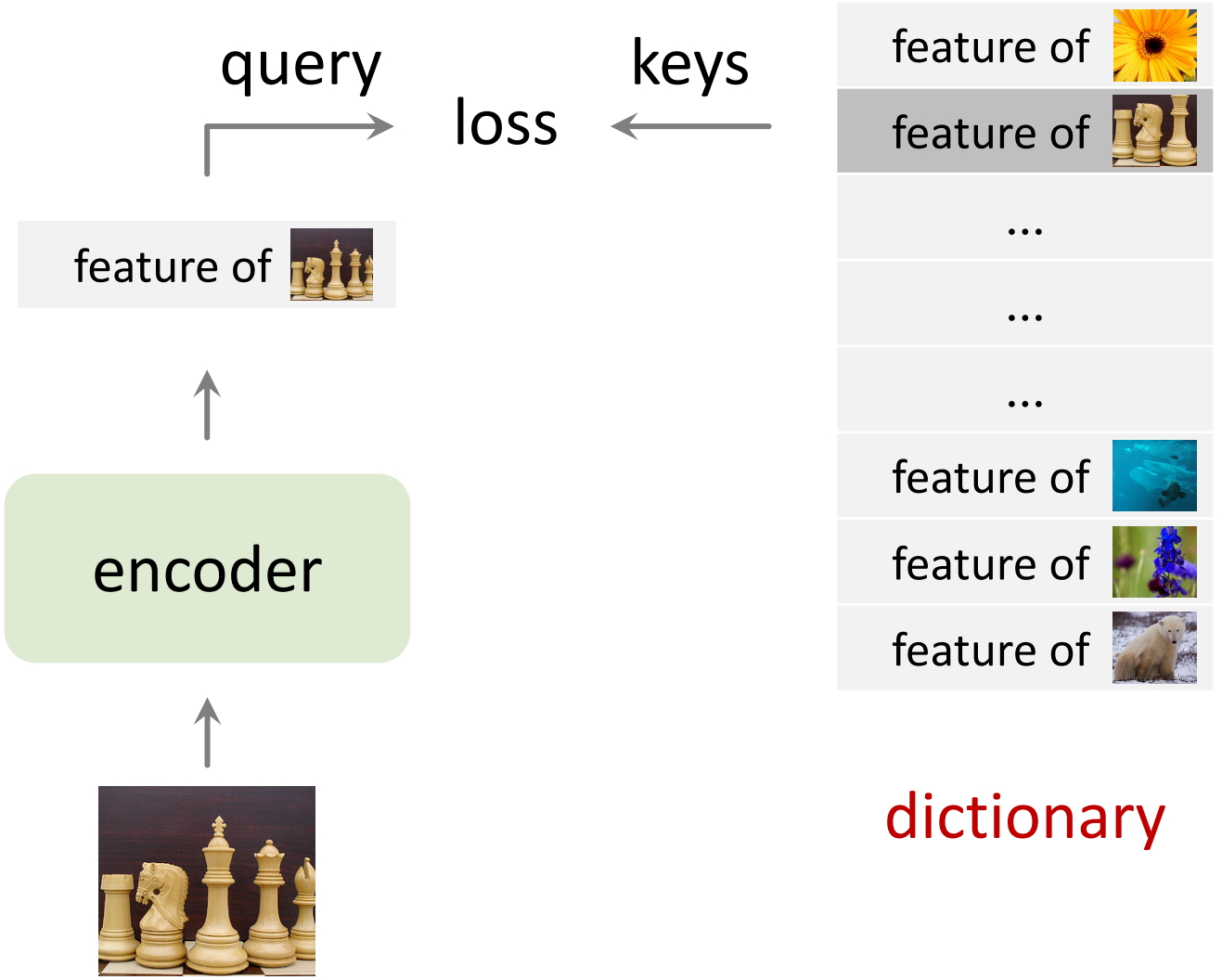
- Unsupervised pre-training: surpass supervised counterparts
- … in 7 vision tasks on detection, segmentation
- … by big margins in some tasks
- … scaled out to 1 billion images

Unsupervised learning in NLP: BERT

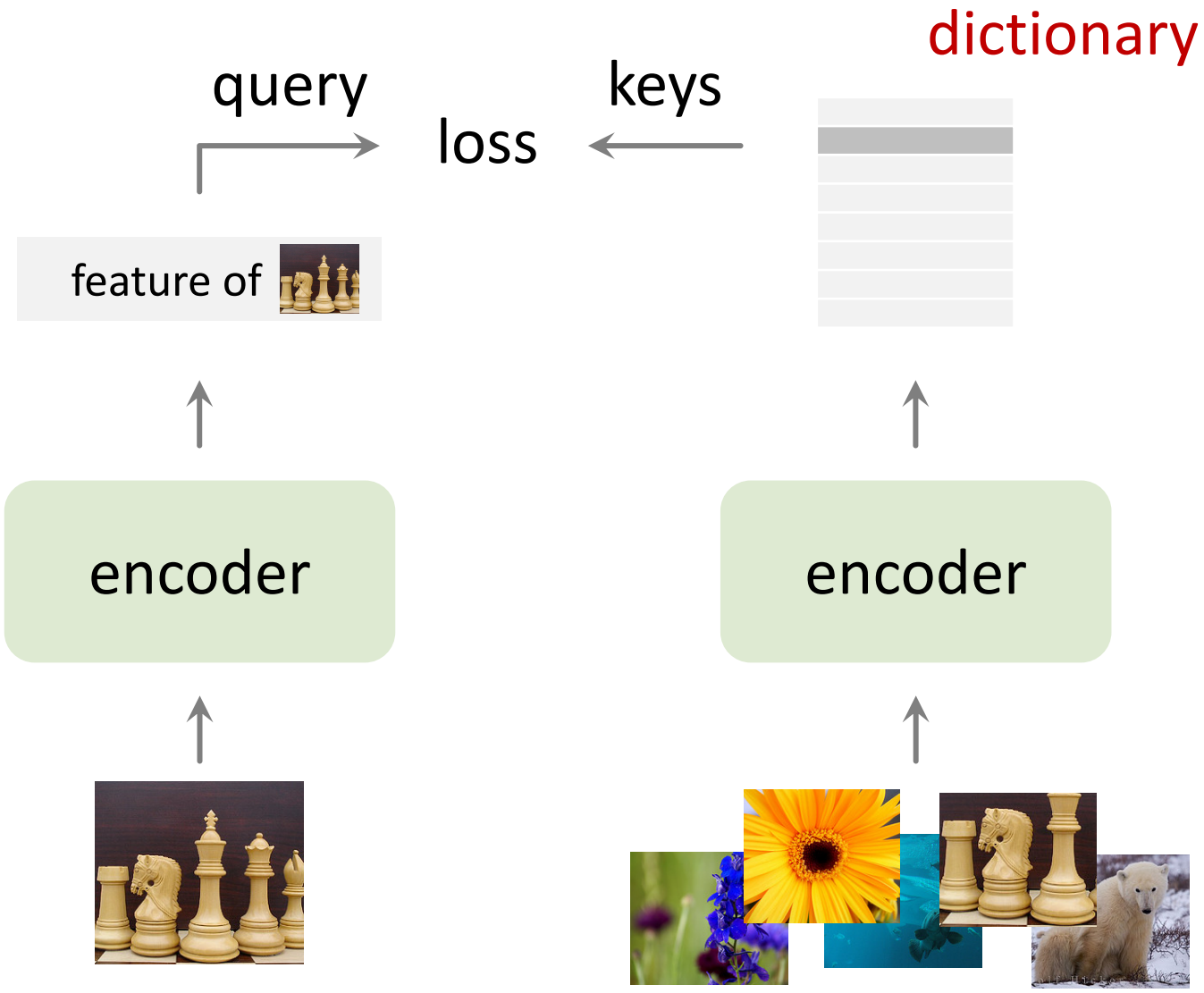


Devlin *et al*. NAACL 2019

Analogy in Computer Vision



Contrastive Learning

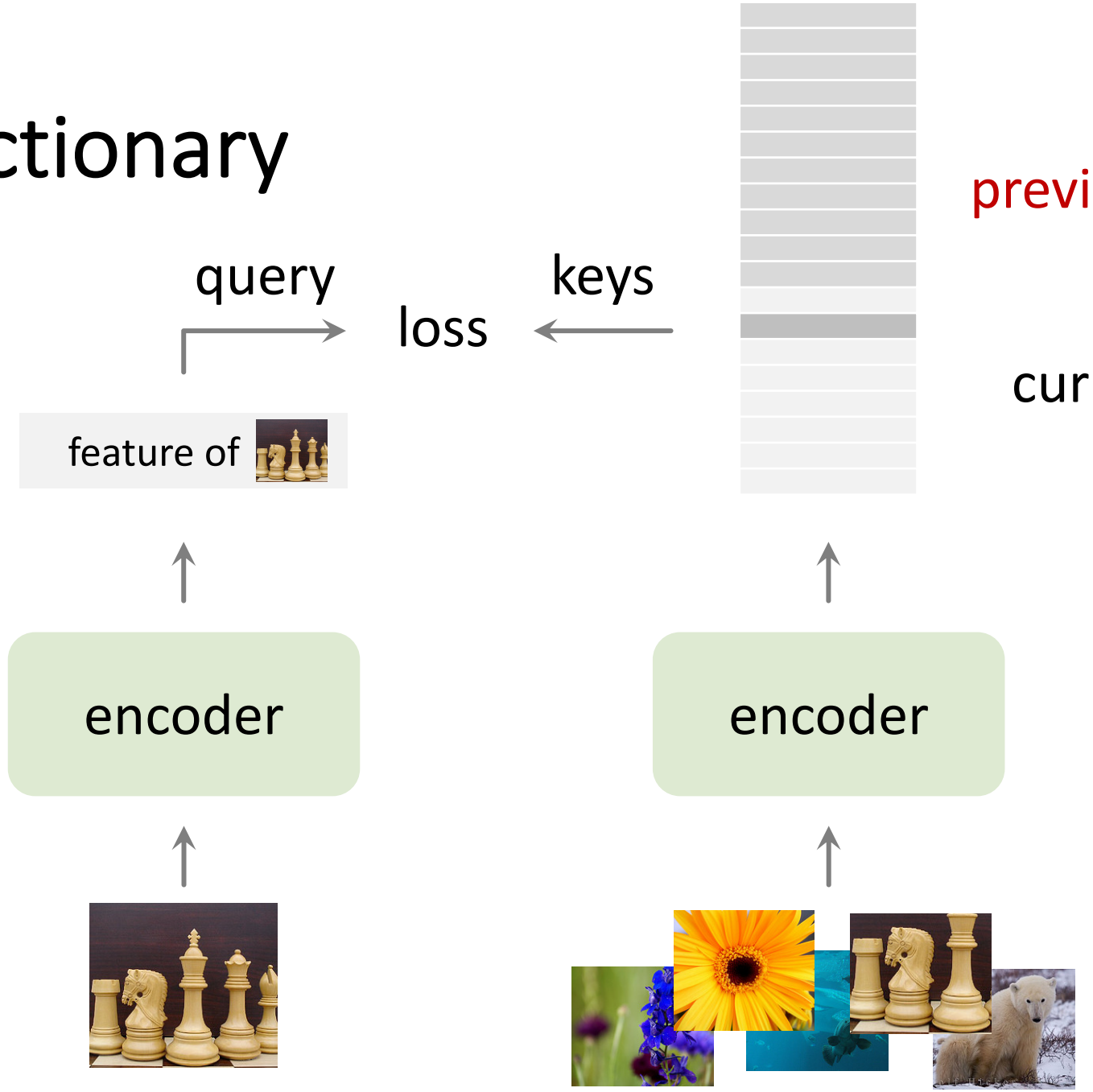


Hadsell *et al*. CVPR 2006,
Wu *et al*. CVPR 2018, …

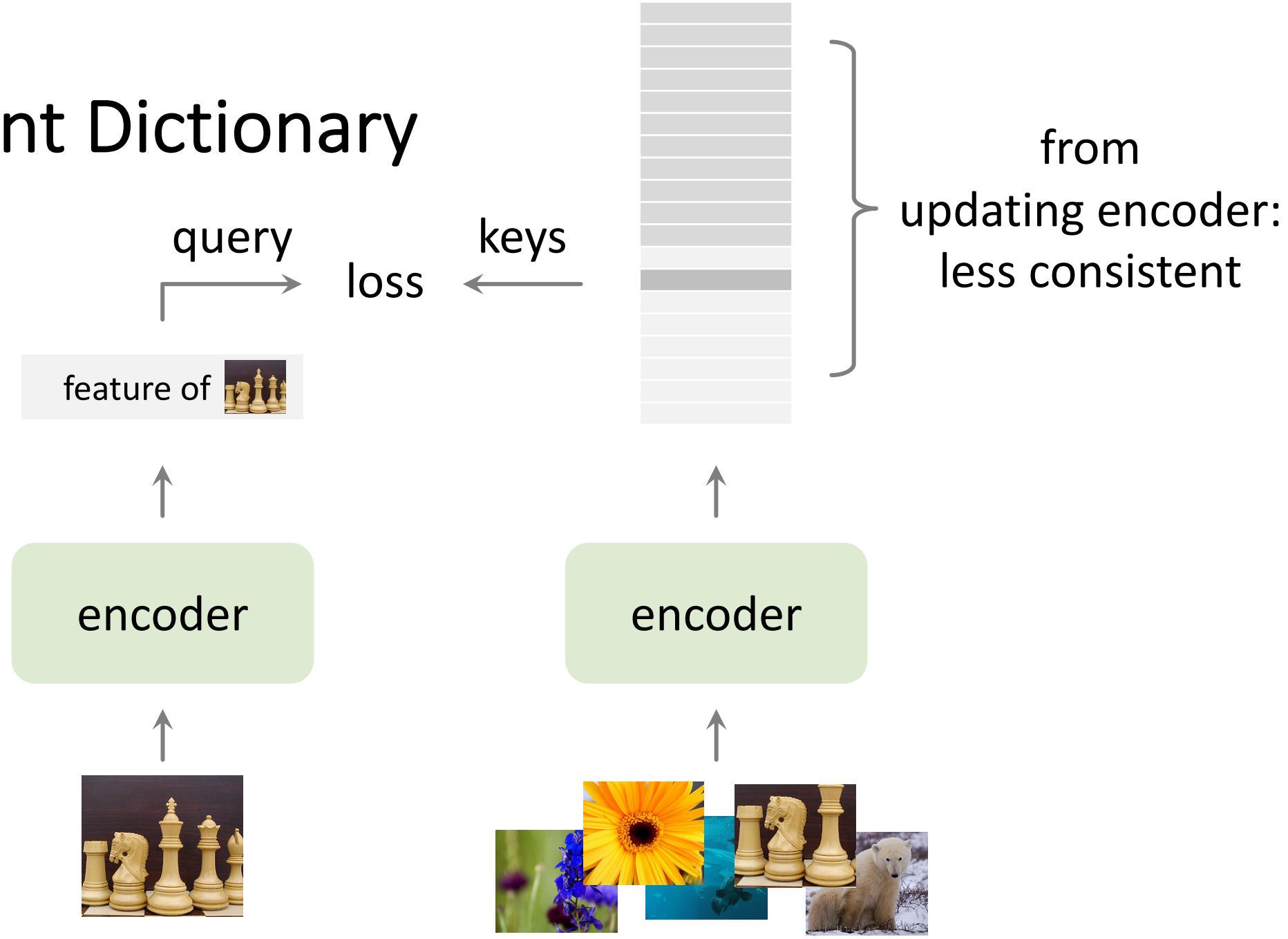
Our Method: Momentum Contrast (MoCo)

- Contrastive learning as dictionary look-up
- **Large** dictionary
- **Consistent** dictionary

Large Dictionary



Consistent Dictionary



Consistent Dictionary

query

keys

loss

from
updating encoder

feature of 

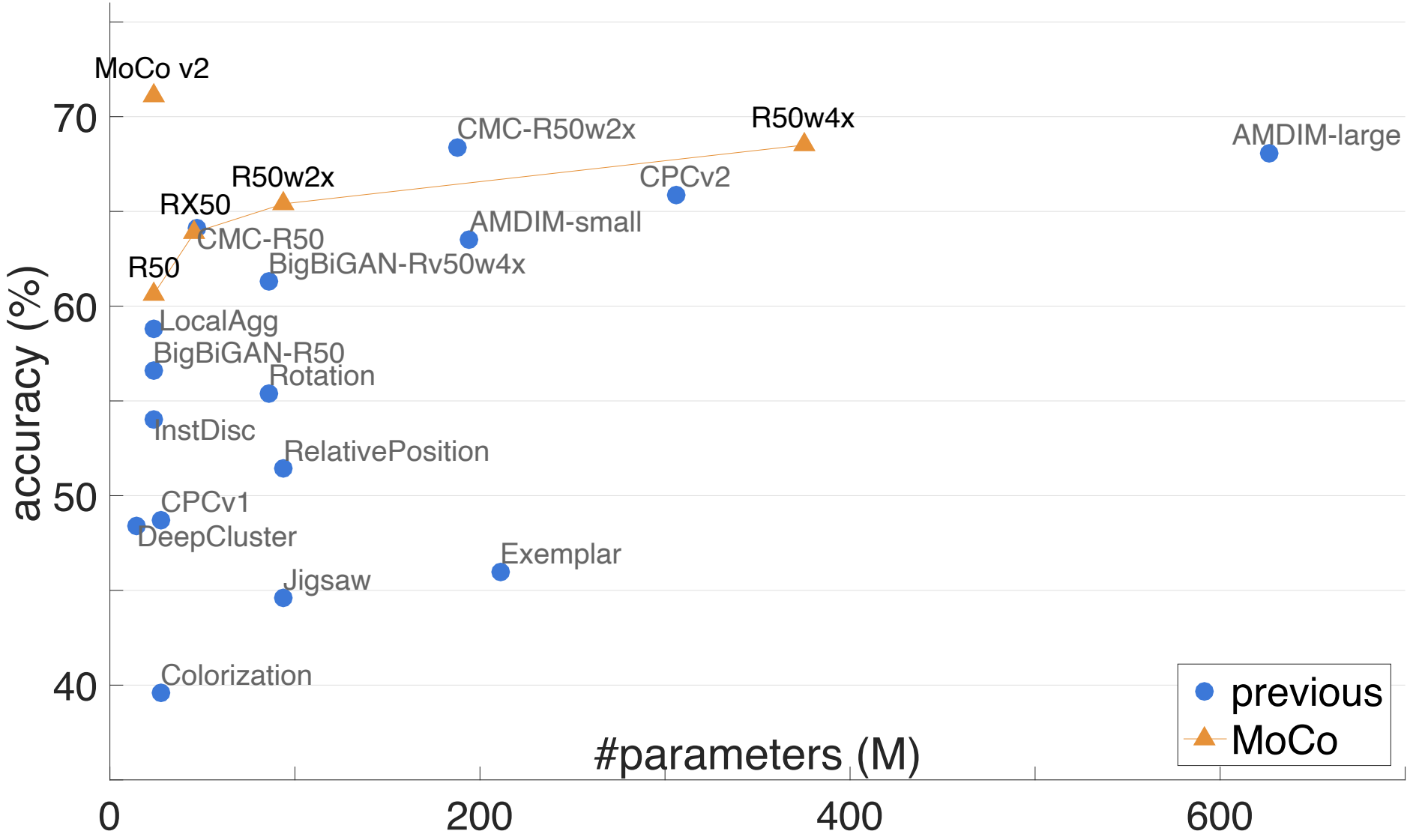
encoder

momentum
encoder

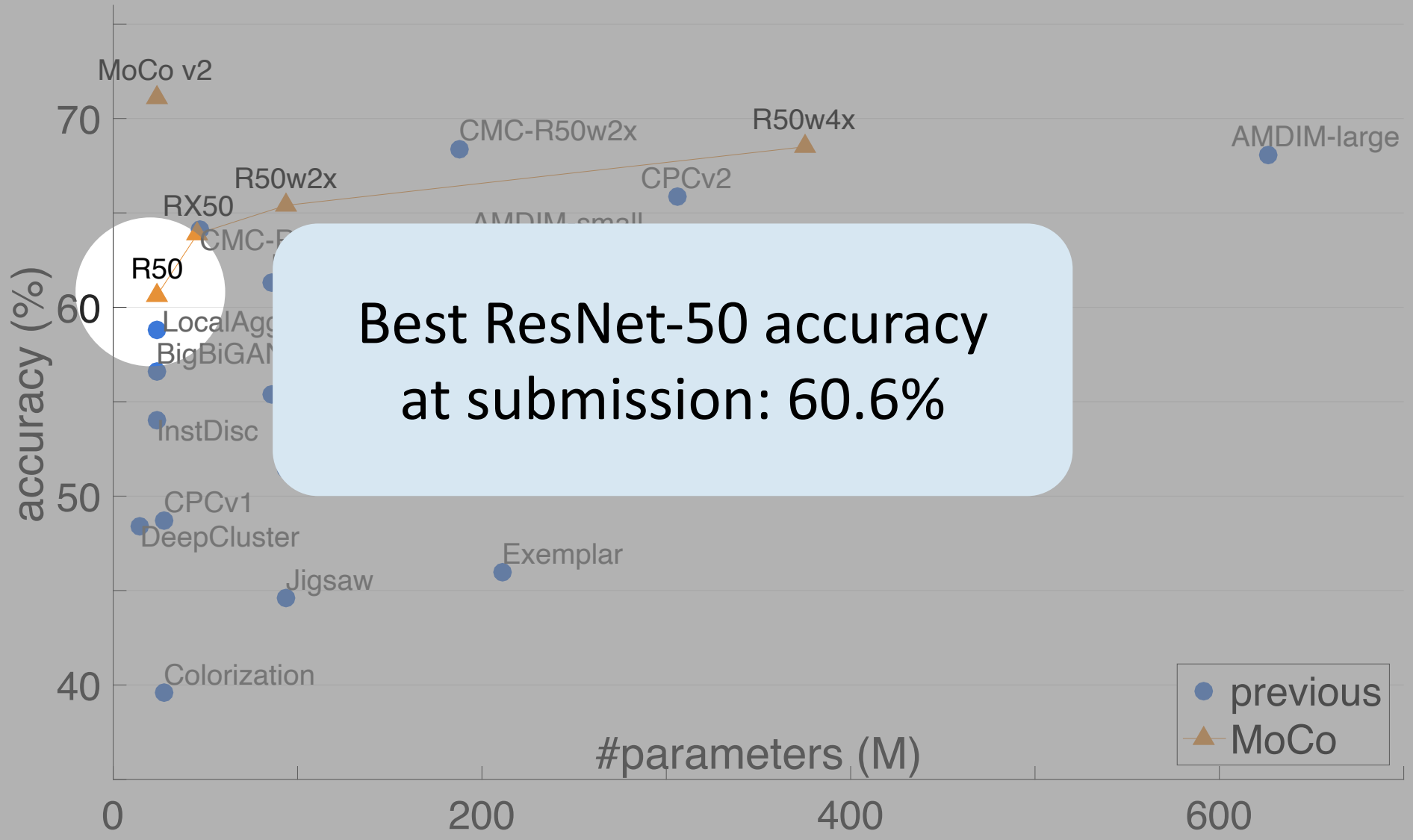
slowly update

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

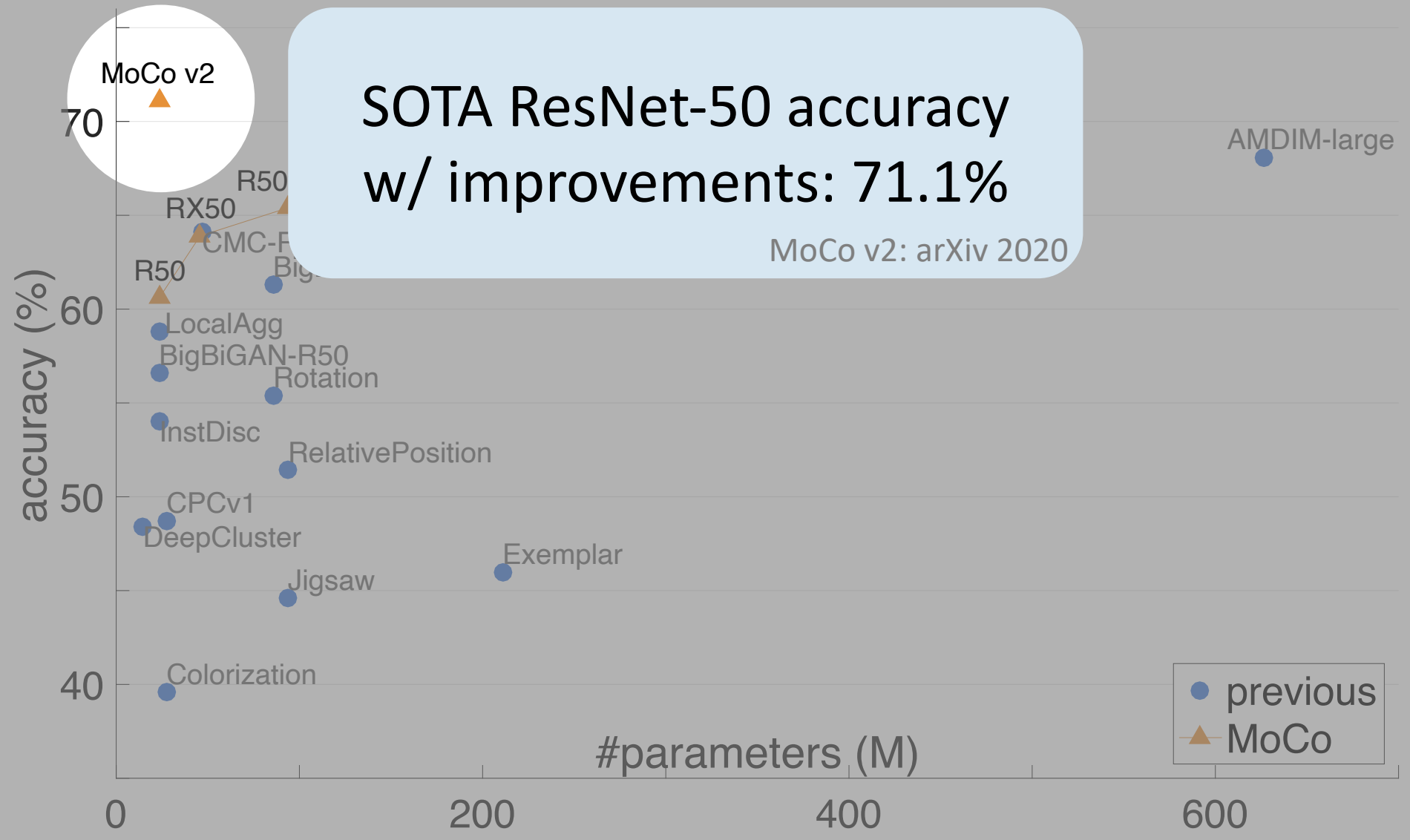
Results: ImageNet Linear Classifiers



Results: ImageNet Linear Classifiers



Results: ImageNet Linear Classifiers



Results: Transferring Features

VOC 2007 Detection, Faster R-CNN, ResNet-50

pre-train	AP_{50}			
	RelPos, by [10]	Multi-task [10]	Jigsaw, by [22]	LocalAgg [60]
super. IN-1M	74.2	74.2	70.5	74.6
unsup. IN-1M	66.8 (−7.4)	70.5 (−3.7)	61.4 (−9.1)	69.1 (−5.5)

Previous:
behind supervised pre-training

Results: Transferring Features

VOC 2007 Detection, Faster R-CNN, ResNet-50

pre-train	AP_{50}				
	RelPos, by [10]	Multi-task [10]	Jigsaw, by [22]	LocalAgg [60]	**MoCo**
super. IN-1M	74.2	74.2	70.5	74.6	74.4
unsup. IN-1M	66.8 (−7.4)	70.5 (−3.7)	61.4 (−9.1)	69.1 (−5.5)	74.9 (+**0.5**)

MoCo:
surpass supervised pre-training

Results: Transferring Features

VOC 2007 Detection, Faster R-CNN, ResNet-50

pre-train	AP_{50}				
	RelPos, by [10]	Multi-task [10]	Jigsaw, by [22]	LocalAgg [60]	**MoCo**
super. IN-1M	74.2	74.2	70.5	74.6	74.4
unsup. IN-1M	66.8 (−7.4)	70.5 (−3.7)	61.4 (−9.1)	69.1 (−5.5)	74.9 (+**0.5**)
unsup. IN-14M	-	-	69.2 (−1.3)	-	75.2 (+**0.8**)
unsup. IG-1B	-	-	-	-	75.6 (+**1.2**)

MoCo:
benefit from 1 billion images

Results: Transferring Features

VOC 2007 Detection, Faster R-CNN, ResNet-50

	AP_{50}	AP	AP_{75}
pre-train	**MoCo**	**MoCo**	**MoCo**
super. IN-1M	74.4	42.4	42.7
unsup. IN-1M	74.9 (+**0.5**)	46.6 (+**4.2**)	50.1 (+**7.4**)
unsup. IN-14M	75.2 (+**0.8**)	46.9 (+**4.5**)	50.2 (+**7.5**)
unsup. IG-1B	75.6 (+**1.2**)	47.6 (+**5.2**)	51.7 (+**9.0**)

MoCo:
big gains in stringent metrics
+9.0 AP_{75}

Results: Transferring Features

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**

AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
36.7	56.7	40.0	33.7	53.8	35.9
40.6	61.3	44.4	36.8	58.1	39.5
40.8 (+0.2)	61.6 (+0.3)	44.7 (+0.3)	36.9 (+0.1)	58.4 (+0.3)	39.7 (+0.2)
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

AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
35.6	54.6	38.2	31.4	51.5	33.5
40.0	59.9	43.1	34.7	56.5	36.9
40.7 (**+0.7**)	60.5 (**+0.6**)	44.1 (**+1.0**)	35.4 (**+0.7**)	57.3 (**+0.8**)	37.6 (**+0.7**)
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

VOC 07+12 Detection

surpass, +4.9 AP_{75}

COCO Detection

COCO Instance seg.

surpass

Results: Transferring Features

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**)

COCO Keypoint
surpass

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**)

COCO Dense pose
surpass, +3.7 AP_{75}

pre-train	LVIS 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**)

LVIS
Instance seg.
surpass

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

Cityscapes
Semantic seg.
surpass

pre-train	Semantic VOC
random init.	39.5
super. IN-1M	74.4
MoCo IN-1M	72.5 (−1.9)
MoCo IG-1B	73.6 (−0.8)

VOC
Semantic seg.
-0.8 point

Conclusion

- Unsupervised pre-training: surpass supervised counterparts

- Code available: https://github.com/facebookresearch/moco