Bag of Instances Aggregation Boosts Self-supervised Distillation

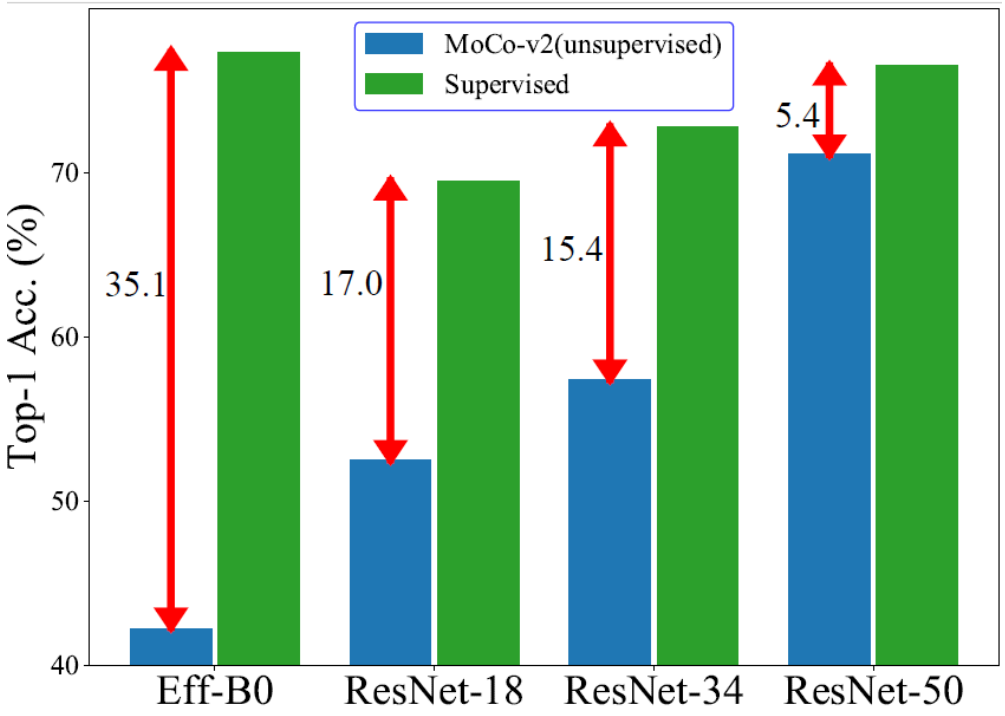
Haohang Xu[1,2*], Jiemin Fang[3,4*], Xiaopeng Zhang[2], Lingxi Xie[2], Xinggang Wang[4], Wenrui Dai[1], Hongkai Xiong[1], Qi Tian[2]

[1]Shanghai Jiao Tong University
[2]Huawei Inc
[3]Institute of Artificial Intelligence, Huazhong University of Science & Technology
[4]School of EIC, Huazhong University of Science & Technology

Code: https://github.com/haohang96/bingo

Knowledge Distillation in Self-supervised Learning

The performance gap between supervised and self-supervised learning (MoCo-v2 as testbed) increases with model's parameters decrease, which is harmful to the deployment on some edge or low computational capacity devices.

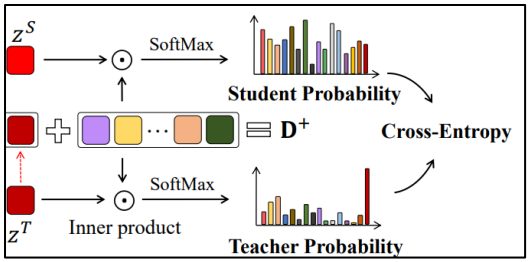


Knowledge Distillation in Self-supervised Learning

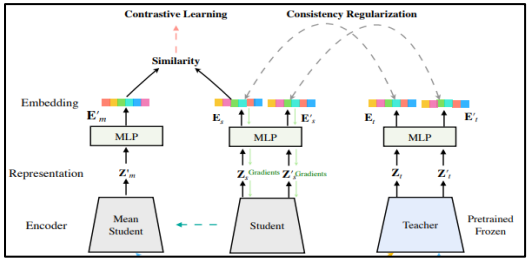
Distillation via **statistics** between the student and teacher

Self-Supervised Distillation

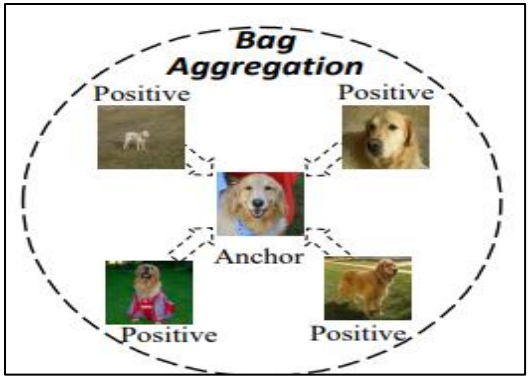
Distributions between randomly selected instances:
SEED, Compress



Embedding distances of one instance's different views:
DisCo, Simreg

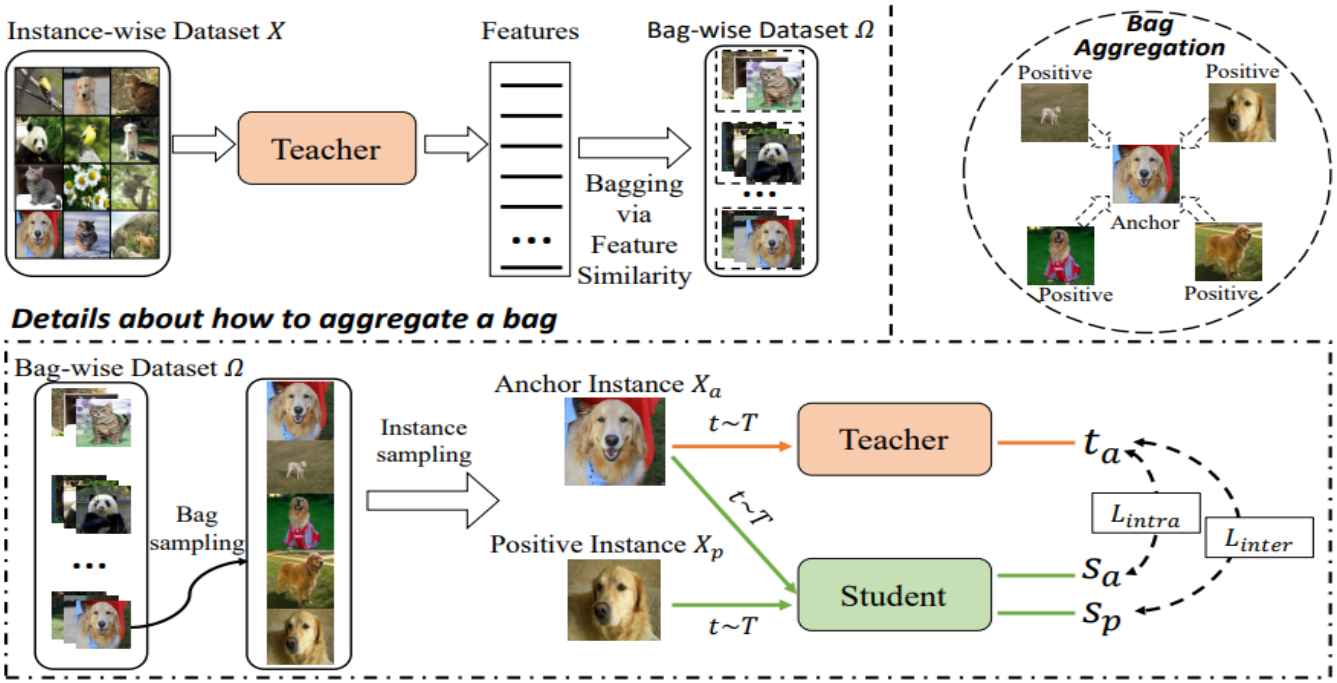


Data relation between similar instances: BINGO (ours)



Knowledge Distillation in Self-supervised Learning

Inter-sample relation Distillation

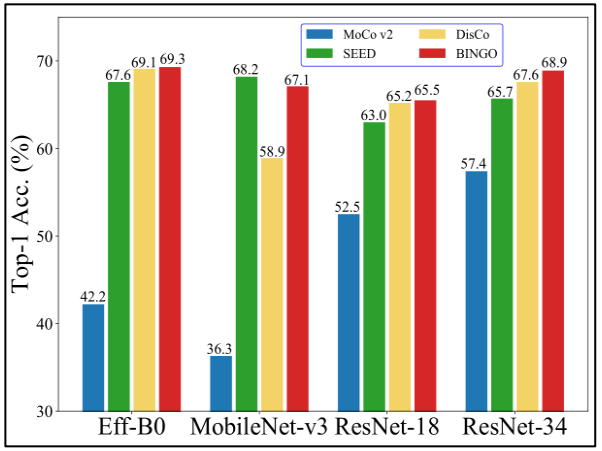


1. Summarize data relation via the teacher network and bag similar samples together
2. Aggregate bags in the bag-wise dataset using L_{inter}

Experiments on ImageNet

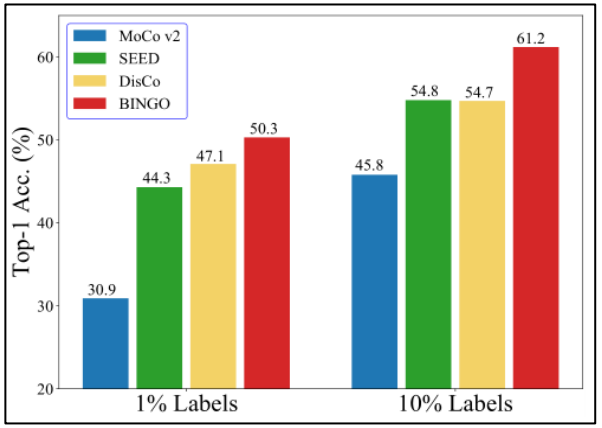
Linear classification

Method	S / T	R-18		R-34	
		T-1	T-5	T-1	T-5
Supervised (Fang et al., 2020)		69.5	-	72.8	-
MoCo-V2 (Baseline) (Fang et al., 2020)		52.5	77.0	57.4	81.6
SEED (Fang et al., 2020)	R-50 (67.4)	57.6	81.8	58.5	82.6
DisCo (Gao et al., 2021)	R-50 (67.4)	60.6	83.7	62.5	85.4
BINGO	R-50 (67.4)	**61.4**	**84.3**	**63.5**	**85.7**
BINGO	R-50 (71.1)	**64.0**	**85.7**	**66.1**	**87.2**
SEED (Fang et al., 2020)	R-152 (74.1)	59.5	83.3	62.7	85.8
DisCo (Gao et al., 2021)	R-152 (74.1)	65.5	86.7	68.1	88.6
BINGO	R-152 (74.1)	**65.9**	**87.1**	**69.1**	**88.9**
SEED (Fang et al., 2020)	R50×2 (77.3)	63.0	84.9	65.7	86.8
DisCo (Gao et al., 2021)	R50×2 (77.3)	65.2	86.8	67.6	88.6
BINGO	R50×2 (77.3)	**65.5**	**87.0**	**68.9**	**89.0**



Semi-supervised Learning

Method	T	1% labels	10% labels
MoCo v2 baseline	-	30.9	45.8
Compress (Abbasi Koohpayegani et al., 2020)	R-50 (67.4)	41.2	47.6
SEED (Fang et al., 2020)	R-50 (67.4)	39.1	50.2
DisCo (Gao et al., 2021)	R-50 (67.4)	39.2	50.1
BINGO	R-50 (67.4)	**42.8**	**57.5**
Compress (Abbasi Koohpayegani et al., 2020)	R-152 (74.1)	-	-
SEED (Fang et al., 2020)	R-152 (74.1)	44.3	54.8
DisCo (Gao et al., 2021)	R-152 (74.1)	47.1	54.7
BINGO	R-152 (74.1)	**50.3**	**61.2**
BINGO	R-50×2 (77.3)	**48.2**	**60.2**



Experiments on downstream tasks

CIFAR-10/100 classification

Method	T	CIFAR-10/100
MoCo v2 baseline	-	77.9/48.1
SEED	R-50 (67.4)	82.3/56.8
DisCo	R-50 (67.4)	85.3/63.3
BINGO	R-50 (67.4)	**86.8/66.5**

COCO object detection and instance segmentation

Method	Mask R-CNN, ResNet-18, Detection												Mask R-CNN, ResNet-18, Instance Segmentation											
	1× schedule						2× schedule						1× schedule						2× schedule					
	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP_S	AP_M	AP_L	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP_S	AP_M	AP_L	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}	AP_S	AP_M	AP_L	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}	AP_S	AP_M	AP_L
MoCo v2	31.3	50.0	33.5	16.5	33.1	41.1	34.4	53.9	37.0	18.9	36.8	45.5	28.8	47.2	30.6	12.2	29.7	42.7	31.5	51.1	33.6	14.1	32.9	46.9
BINGO	**32.0**	**51.0**	**34.7**	**17.1**	**34.1**	**42.0**	**34.9**	**54.2**	**37.7**	**20.0**	**37.1**	**46.0**	**29.6**	**48.2**	**31.5**	**12.8**	**30.8**	**43.0**	**31.9**	**51.7**	**33.9**	**14.9**	**33.1**	**47.2**

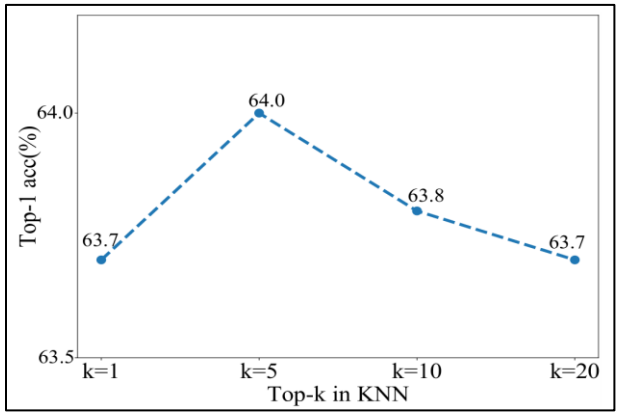
Method	Mask R-CNN, ResNet-34											
	Object Detection						Instance Discrimination					
	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP_S	AP_M	AP_L	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}	AP_S	AP_M	AP_L
MoCo v2	38.1	56.8	40.7	-	-	-	33.0	53.2	35.3	-	-	-
SEED Fang et al. (2020)	38.4	57.0	41.0	-	-	-	33.3	53.7	35.3	-	-	-
DisCo Gao et al. (2021)	39.4	58.7	42.7	-	-	-	34.4	55.4	36.7	-	-	-
BINGO	**39.9**	**59.4**	**43.5**	22.8	43.3	52.1	**35.7**	**56.5**	**38.2**	16.8	37.9	51.6

Effects of bagging strategy

Effects of teacher parameters and relation

Teacher Parameters	Student Relation	Teacher Relation	Accuracy
✗	✗	✗	52.2 (w/o distillation)
✗	✓	✗	57.2
✗	✗	✓	62.2
✓	✗	✗	62.0
✓	✓	✗	62.5
✓	✗	✓	64.0

Effect of K in K-nearest neighbors



Using K-means as bagging strategy

Number of Clusters (C)	Accuracy
5000	62.7
10000	63.5
20000	63.8
50000	63.6