ISD: Self-Supervised Learning by Iterative Similarity Distillation

Ajinkya Tejankar[1,*] Soroush Abbasi Koohpayegani[1,*] Vipin Pillai[1] Paolo Favaro[2] Hamed Pirsiavash[3]

[1]University of Maryland Baltimore County University Of Bern [3]University of California, Davis

* Equal contribution

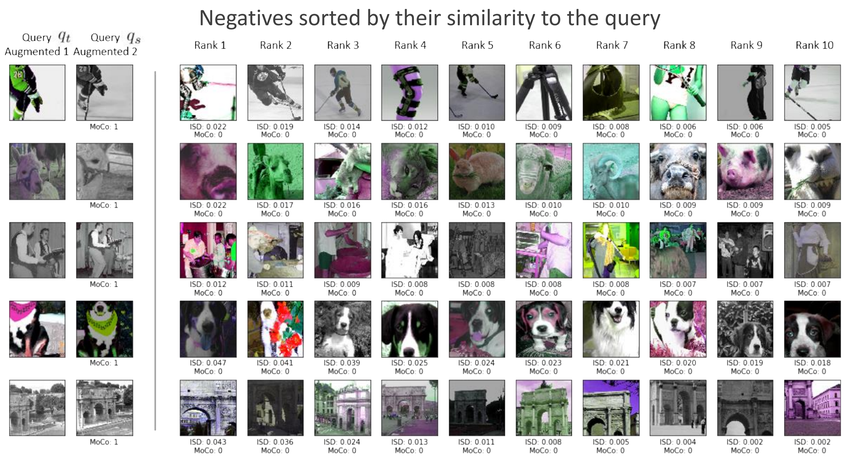
2021 ICCV OCTOBER 11-17 VIRTUAL

PyTorch Code

https://github.com/UMBCvision/ISD

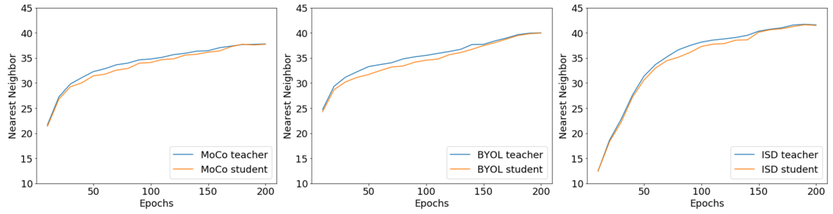
Motivation

- **Goal:** Learning rich representations from unlabeled images
- **Challenge:** Some negative samples in contrastive learning (e.g., MoCo) may be semantically related to the positive one, so may be misleading.



Iterative Similarity Distillation

- We can view the recent direction of SSL methods as iterative self-distillation where there is a teacher and a student.
- Our study shows that the teacher is always better in intermediate checkpoints
- We iteratively distill the teacher to the student and update the teacher as a moving average



Key Ideas

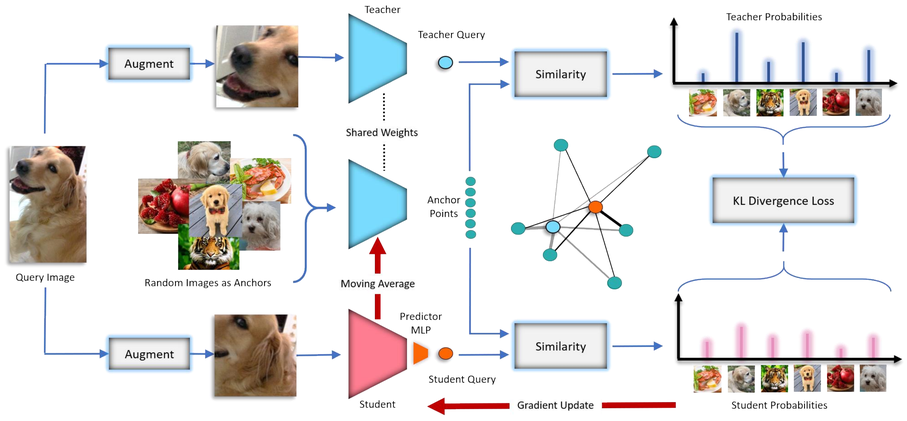
- In contrastive learning, all negatives are not equally negative
  - Relaxing **binary** instance classification of contrastive learning with **soft** classification
- Since teacher has better accuracy, use the teacher as the source of knowledge to assign soft-labels to negatives

Method

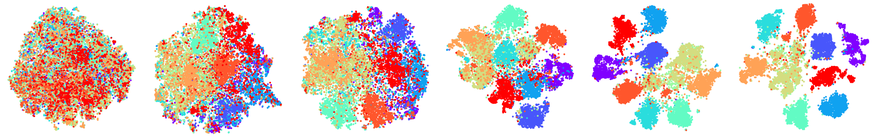
1. Augment an image **twice**
2. Capture the **similarity** of the **first** view to anchors in the teacher's embedding space.
3. Capture the **similarity** of the **second** view to anchors in the student's space.
4. Train the student by minimizing the **difference in similarities** (KL divergence).
5. Update the teacher as a **moving average** of the student.

$$\theta_t \leftarrow m\theta_t + (1-m)\theta_s$$

Teacher Parameters — Momentum Value — Student Parameters



t-SNE of different checkpoints in our method for 10 class of ImageNet



Results

On ImageNet task, our method outperforms other SOTA methods in ResNet50 and ResNet18.

Method	Batch Size	Epochs	Sym. Loss 2x FLOPS	Top-1 Linear	NN	20-NN
ResNet-50						
Supervised	256	100	-	76.2	71.4	74.8
SwAV	4096	200	✓	69.1	-	-
SimCLR	4096	1000	✓	69.3	-	-
MoCo-V2	256	200	✓	69.9	-	-
SimSiam	256	200	✓	70.0	-	-
BYOL	4096	200	✓	70.6	-	-
MoCo-V2	256	400	✓	71.0	-	-
MoCo-V2	256	800	✗	71.1	57.3	61.0
CompRess	256	1K+130	✗	71.9	63.3	66.8
BYOL	4096	1000	✓	74.3	62.8	66.9
SwAV	4096	800	✓	**75.3**	-	-
MoCo-V2	256	200	✗	67.5	-	-
CO2	256	200	✗	68.0	-	-
BYOL-asym	256	200	✗	69.3	55.0	59.2
ISD	256	200	✗	**69.8**	**59.2**	**62.0**
ResNet-18						
Supervised	256	100	-	69.8	63.0	67.6
MoCo-V2	256	200	✗	51.0	37.7	42.1
BYOL-asym	256	200	✗	52.6	40.0	44.8
ISD	256	200	✗	**53.8**	**41.5**	**46.6**

Comparison on ImageNet with Limited Labels for ResNet-50

Method	Epochs	Top-1 1%	Top-1 10%	Top-5 1%	Top-5 10%
Entire network is fine-tuned.					
Supervised		25.4	56.4	48.4	80.4
PIRL	800	-	-	57.2	83.8
CO2	200	-	-	71.0	85.7
SimCLR	1000	48.3	65.6	75.5	87.8
InvP	800	-	-	78.2	88.7
BYOL	1000	53.2	68.8	78.4	89.0
SwAV	800	**53.9**	**70.2**	**78.5**	**89.9**
Only the linear layer is trained.					
BYOL	1000	55.7	**68.6**	80.0	**88.6**
CompRess	1K+130	**59.7**	67.0	**82.3**	87.5
MoCo v2	200	43.6	58.4	71.2	82.9
BYOL-asym	200	47.9	61.3	74.6	84.7
ISD	200	**53.4**	**63.0**	**78.8**	**85.9**