Mean Shift for Self-Supervised Learning

Soroush Abbasi Koohpayegani[1,*] Ajinkya Tejankar[1,*] Hamed Pirsiavash[2]
[1]University of Maryland Baltimore County [2]University of California, Davis
* Equal contribution

2021 ICCV OCTOBER 11-17 VIRTUAL

PyTorch Code

https://github.com/UMBCvision/MSF

Motivation

- **Goal**: Learning rich representations from unlabeled images
- Two recent directions of Self-Supervised Learning:
  1. Pull two augmentation of the same image closer (e.g., MoCo, BYOL)
  2. Cluster similar images together (e.g., SwAV, DeepCluster, SeLA)
- We are interested in an SSL method that groups similar images together without:
  - explicit clustering
  - contrasting between data points or clusters

Key Idea

- Inspired by mean-shift algorithm, we pull a data point closer toward not only its other augmentations but also its nearest neighbors (NNs)
- No explicit clustering, so no priors on the shape, size, of number of the clusters
- No contrast between different images or even cluster centers

Method

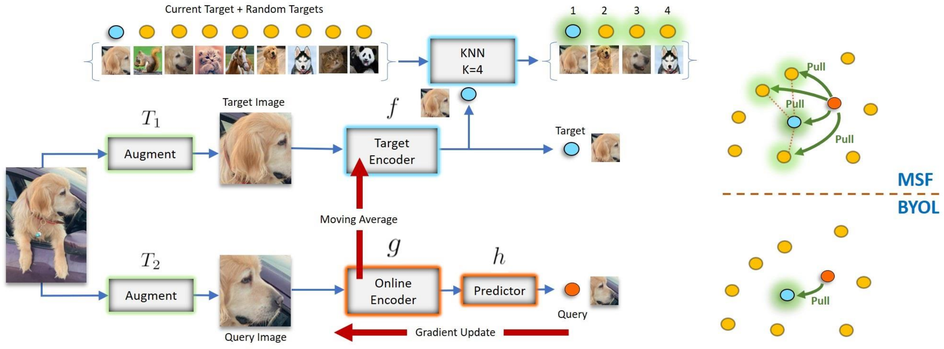
- Augment image x twice: $T_1(x)$ and $T_2(x)$
- Forward them through both Target and Online encoders :

$$u = f(T_1(x)) \qquad v = h(g(T_2(x)))$$

Target Encoder; Prediction Head; Online Encoder

- Let $\{z_j\}_{j=1}^{k}$ be k-nearest neighbors of u in a set that includes u
- Optimize the Online Encoder using following loss:
- We use Cosine distance for $dist(.)$ and kNN

$$L = \frac{1}{k}\sum_{j=1}^{k} dist(v, z_j)$$



1 Forward tne first view through the Target Encoder to get the Target

2 Forward the second view through the Online Encoder followed by Prediction Head to get the Query

3 Get the k-nearest neighbors of the Target

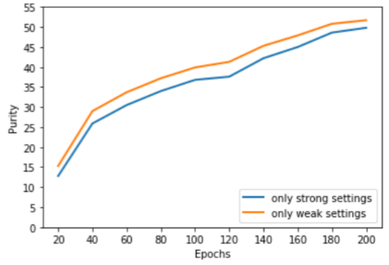
4 Mean-Shift: Pull the Query to be close to Target and its nearest neighbors

5 Optimize the loss and update the Online Encoder using gradient decent.

6 Update the Teacher Encoder as the moving average of the Online Encoder.

Augmentation Strategies

- Two augmentation options for training:
  - **Weak** Augmentation: Random Resize Crop + Random Horizontal Flip
  - **Strong** Augmentation: Weak augmentation + Random {Blur , Color Jitter , Gray Scale}
- **Our observation:** Strong augmentation makes target image unnatural, leading to noisy kNN results
  - **Purity** is defined as: percentage of kNNs with the same label as the target
  - **MSF(W/S):** Use weak augmentation for target and strong augmentation for online encoders

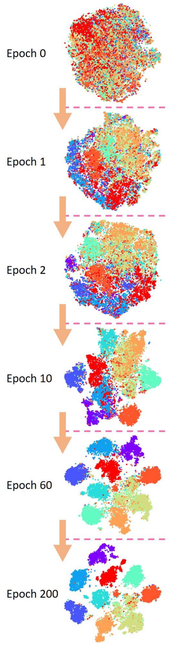


Method	Aug.	Top-1	NN	20-NN
BYOL-asym	s/s	69.3	55.0	59.2
BYOL-asym	w/s	69.5	55.8	59.1
BYOL	w/w	60.1	-	-
MSF	s/s	71.4	60.6	64.0
MSF	w/s	72.4	62.0	64.9
MSF	w/w	66.3	54.6	57.4

Results

- Our method outperforms SOTA with more than 2 points in ResNet50 with similar training computation budget (256 batch size, 200 epoch)
- It outperform BYOL with more than 6 points when using weak augmentation only
  - Suitable for applications with no obvious augmentation options, e.g., in medical domain
- t-SNE of different checkpoints in our method for 10 class of ImageNet
  - Our method learns to cluster semantically similar images together.

Method	Batch Size	Epochs	Sym. Loss 2x FLOPS	Top-1 Linear	NN	20-NN
Supervised	256	100	-	76.2	71.4	74.8
Random-init	-	-	-	5.1	1.5	2.0
SeLa-v2	4096	400	✓	67.2	-	-
SimCLR	4096	1000	✓	69.3	-	-
SwAV	4096	400	✓	70.1	-	-
DeepCluster-v2	4096	400	✓	70.2	-	-
SimSiam	256	400	✓	70.8	-	-
MoCo v2	256	800	✗	71.1	57.3	61.0
CompRess	256	1K+130	✗	71.9	63.3	66.8
InvP	256	800	✗	71.3	-	-
BYOL	4096	1000	✓	74.3	62.8	66.9
SwAV	4096	800	✓	**75.3**	-	-
SimCLR	4096	200	✓	68.3	-	-
SwAV	4096	200	✓	69.1	-	-
MoCo v2	256	200	✓	69.9	-	-
SimSiam	256	200	✓	70.0	-	-
BYOL	4096	200	✓	**70.6**	-	-
MoCo v2	256	200	✗	67.5	50.9	54.3
CO2	256	200	✗	68.0	-	-
BYOL-asym	256	200	✗	69.3	55.0	59.2
ISD	256	200	✗	69.8	59.2	62.0
MSF	256	200	✗	71.4	60.6	64.0
MSF w/s	256	200	✗	**72.4**	**62.0**	64.9
MSF w/s (128K)	256	200	✗	72.1	**62.0**	**65.2**
SimCLR w/w	4096	300	✓	40.2	-	-
BYOL w/w	4096	300	✓	60.1	-	-
MSF w/w	256	200	✗	**66.3**	54.6	57.4



- Our method outperforms SOTA in transfer learning (10 dataset)

Method	Food 101	CIFAR 10	CIFAR 100	SUN 397	Cars 196	Aircraft	DTD	Pets	Caltech 101	Flowers 102	Mean
Supervised	72.3	93.6	78.3	61.9	66.7	61.0	74.9	91.5	94.5	94.7	78.9
BYOL-asym	70.2	91.5	74.2	59.0	54.0	52.1	**73.4**	86.2	90.4	92.1	74.3
MoCo v2	70.4	91.0	73.5	57.5	47.7	51.2	73.9	81.3	88.7	91.1	72.6
MSF	70.7	92.0	76.1	59.0	**60.9**	53.5	72.1	89.2	92.1	92.4	75.8
MSF-w/s	71.2	92.6	**76.3**	59.2	55.6	53.7	73.2	88.7	**92.7**	92.0	75.5
MSF-w/s (128K)	**72.3**	**92.7**	**76.3**	**60.2**	59.4	**56.3**	71.7	**89.8**	90.9	**93.7**	**76.3**