Constrained Mean Shift Using Distant Yet Related Neighbors for Representation Learning

Navaneet K L[1,*], Soroush Abbasi Koohpayegani[1,*], Ajinkya Tejankar[1,*], Kossar Pourahmadi[1], Akshayvarun Subramanya[2], Hamed Pirsiavash[1,2]

[1]University of California, Davis [2]University of Maryland, Baltimore County

* Equal contribution

ECCV TEL AVIV 2022

PyTorch https://github.com/UCDvision/CMSF

Motivation

Goal: Learn rich representations through self- and semi-supervised learning for image classification.

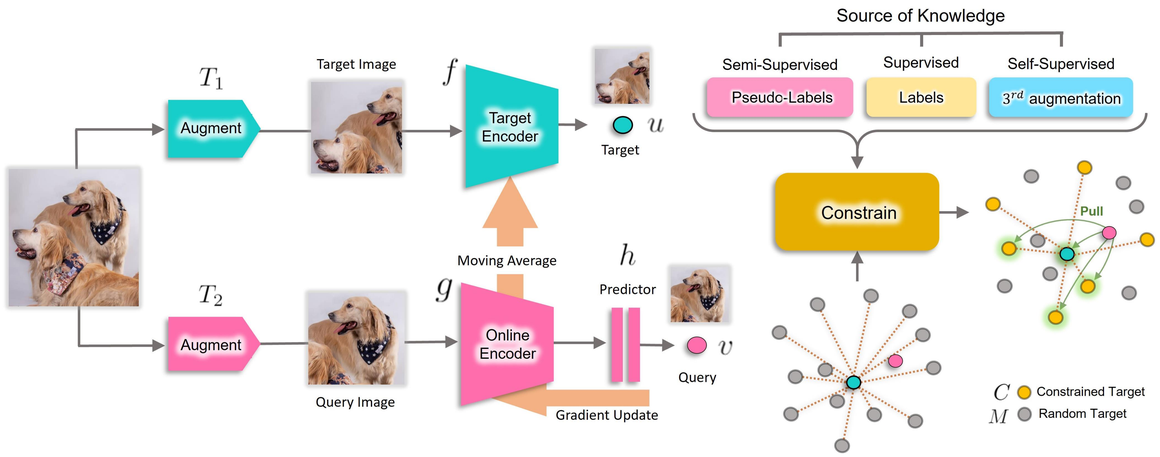
- Recent works [MSF, NNCLR] pull embeddings of images towards their nearest neighbors (NN)
- However, by definition, NNs are close to the original image, so they may not provide a strong training signal.
- We want to find far away image embeddings that are semantically close to the original image.
- **Our idea:** constrain the NN search space using some extra knowledge so that NNs are far
  - **Self-supervised:** Augmentations from previous epoch
  - **Semi-supervised:** Pseudo-labels
  - **Noisy Supervised:** Ground-truth labels
  - **Cross-modal learning:** NNs in other modalities

We show the rank of constrained NNs in unconstrained set. Nearby samples in unconstrained set (rank-1) are not semantically related.

Method

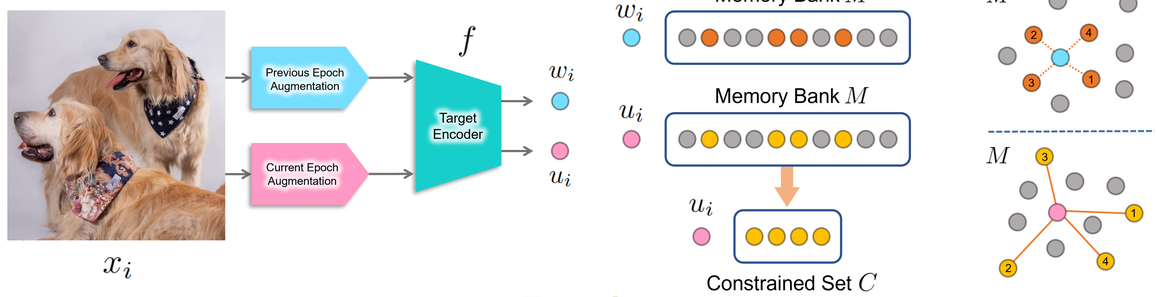
Semi- and Noisy Supervised:

- Define constrained set C as only those images that share the same ground-truth or pseudo-label as the original image
- Since we still search for NNs in C, we are robust to noise in set C
- It better preserves latent structure of categories.



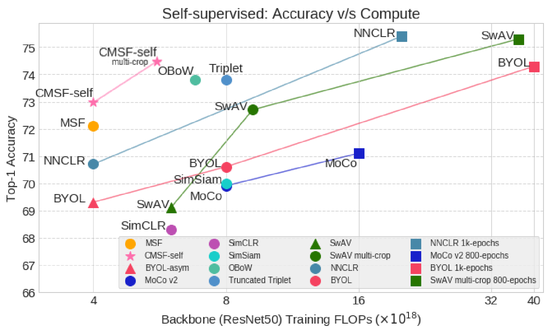
Self-supervised:

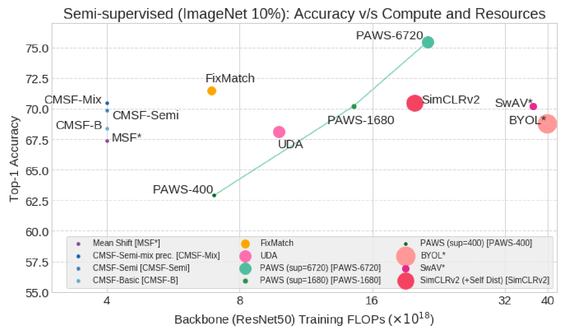
- First, find NNs of another augmentation, use corresponding new augmentations to define set C.
- Second, limit NN search space to images in set C only.
- For efficient implementation, cache the augmentation from the previous epoch in a new memory bank M'.



Results

- Compute and resource (GPUs) efficient.
- Outperforms all methods with similar compute on both self- and semi-supervised setups.

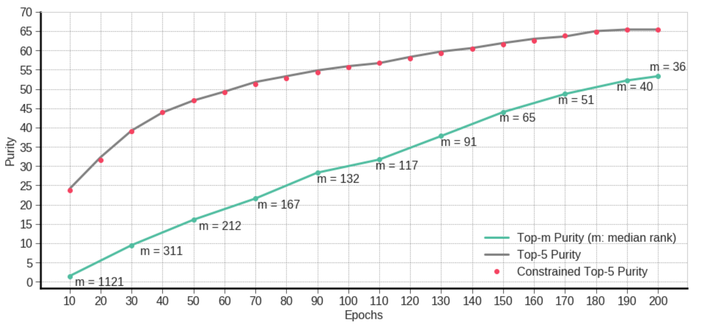




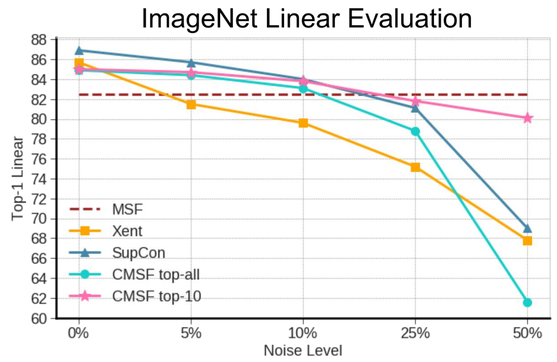
Results

Method	Ref.	Batch Size	Epochs	Sym. Loss 2x FLOPS	Multi-Crop Training	Top-1 Linear	NN	20-NN
NNCLR[21]	[21]	4096	200	✗	✗	70.7	-	-
BYOL [27]	[16]	4096	200	✓	✗	70.6	-	-
SwAV [11]	[16]	256	200	✓	✓	72.7	-	-
Truncated Triplet [67]	[67]	832	200	✓	✗	73.8	-	-
OBoW [24]	[24]	256	200	✗	✓	73.8	-	-
$CMSF_{self}$ (128K)	-	256	200	✗	✓	**74.4**	**62.3**	**66.2**
MSF (1M) [37]	[37]	256	200	✗	✗	72.4	62.0	64.9
MSF (256K)[37]	[37]	256	200	✗	✗	72.2	62.1	65.1
$CMSF_{self}$ (128K)	-	256	200	✗	✗	**73.0**	**63.2**	**66.4**

- Simply increasing number of NNs reduces purity
- While our method preserves high purity with far away NNs



- Relatively robust to label noise in supervised setting



References:

[MSF] Koohpayegani *et al.* Mean shift for self-supervised learning. *ICCV'21*

[NNCLR] Dwibedi *et al.* With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. *ICCV'21*