Incremental False Negative Detection for Contrastive Learning

Tsai-Shien Chen[1], Wei-Chih Hung[2], Hung-Yu Tseng[3], Shao-Yi Chien[1], Ming-Hsuan Yang[3,4,5]

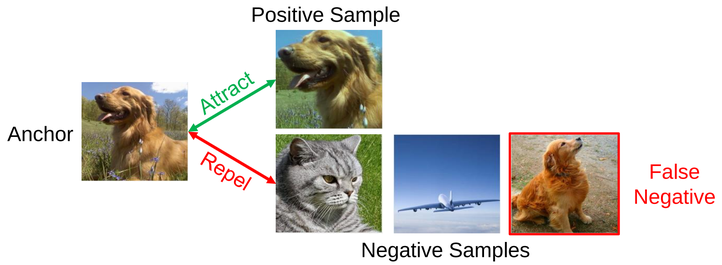
[1]National Taiwan University, [2]Waymo LLC, [3]University of California, Merced, [4]Yonsei University, [5]Google Research

ICLR

10th International Conference on Learning Representations 2022

Problem Statement

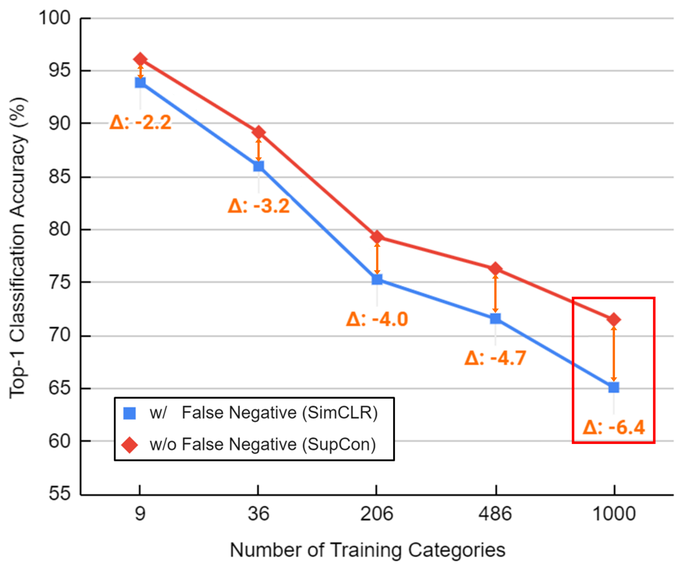
- **False Negative:** the negative sample which shares a similar semantic meaning with the anchor.



- Training with false negatives will adversely affect the self-supervised contrastive learning. [Saunshi et al., 2019]

Effect of False Negative Samples

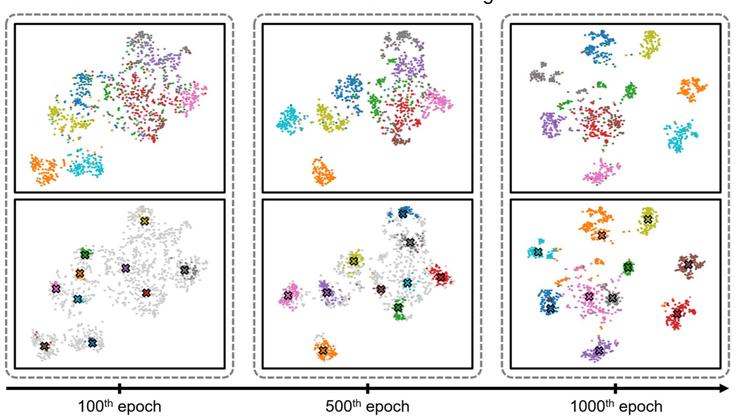
- To analyze the effects of false negatives, we compare two frameworks:
- **SimCLR** [Chen et al., 2020]: Instance-level contrastive learning that trains **with** false negatives.
- **SupCon** [Khosla et al., 2020]: Supervised contrastive learning that trains **without** false negatives.



- Experiments shows that for the **datasets with more classes**, there are **larger performance drops** due to the training with false negatives.

Methodology

- Part 1: How to **detect** false negatives?
- The images with the same pseudo label (assigned by clustering) as the anchor are detected as false negatives.
- We propose a strategy that **uses pseudo labels in an incremental way**.



- In the early, we only use a few labels and the learning is still like **instance-level contrastive learning**.
- In the later, we use more high-quality pseudo labels to benefit from **semantic-aware representation learning**.

- Part 2: How to **remove** False Negative Samples?
- Instance-level contrastive loss is denoted as:

$$\mathcal{L}_{inst} = \sum_{i \in \mathcal{I}} -\log \frac{\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_{i'})}{\sum_{s \in \mathcal{S}(i)} \text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_s)}, \quad \mathcal{S}(i) \equiv \{i', n \mid n \in \mathcal{N}(i)\}$$

- We discuss two losses to remove the detected false negatives:
- **Elimination loss** directly eliminates the false negatives from training batch:

$$\mathcal{L}_{elim} = \sum_{i \in \mathcal{I}} -\log \frac{\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_{i'})}{\sum_{s \in \mathcal{S}(i)} \text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_s)}, \quad \mathcal{S}(i) \equiv \{i', n \mid n \in \mathcal{N}(i), y_n \neq y_i\}$$

- **Attraction loss** treats the detected false negatives as positive samples:

$$\mathcal{L}_{attr} = \sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} -\log \frac{\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_p)}{\sum_{s \in \mathcal{S}(i)} \text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_s)},$$

$$\begin{cases} \mathcal{S}(i) \equiv \{i', n \mid n \in \mathcal{N}(i)\} \\ \mathcal{P}(i) \equiv \{i', n \mid n \in \mathcal{N}(i), y_n = y_i\} \end{cases}$$

- Both the theoretical and quantitative analysis show that attraction loss is more sensitive to noisy pseudo labels and unsuitable for self-supervised learning.

Experiments

- Linear evaluation and transfer learning on three benchmarks:

Method	Architecture	Pre-training batchsize	Pre-training epochs	Datasets ImageNet	Datasets VOC	Datasets Places
SimCLR (Chen et al., 2020b)	ResNet-50	256	200	64.3	-	-
MoCo (He et al., 2020)	ResNet-50	256	200	60.6	79.2	48.9
MoCo v2 (Chen et al., 2020d)	ResNet-50	256	200	67.5	84.0	50.1
PCL (Li et al., 2021)	ResNet-50	256	200	67.6	85.4	50.3
IFND (Ours)	ResNet-50	256	200	**69.7**	**87.3**	**51.9**
SimCLR (Chen et al., 2020b)	ResNet-50	4096	1000	69.3	-	-
BYOL (Grill et al., 2020)	ResNet-50	4096	1000	74.3	-	-
SwAV (Caron et al., 2020)	ResNet-50	4096	800	75.3	88.9	56.7

- Semi-supervised learning on ImageNet:

Method	Architecture	Pre-training batchsize	Pre-training epochs	Label fraction 1%	Label fraction 10%
MoCo (He et al., 2020)	ResNet-50	256	200	56.9	83.0
MoCo v2 (Chen et al., 2020d)	ResNet-50	256	200	66.3	84.4
PCL (Li et al., 2021)	ResNet-50	256	200	75.3	85.6
IFND (Ours)	ResNet-50	256	200	**77.0**	**86.5**

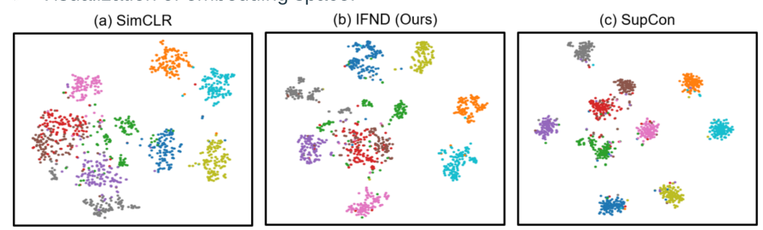
- Object detection and instance segmentation on COCO:

Method	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
Supervise	40.0	59.9	43.1	34.7	56.5	36.9
MoCo (He et al., 2020)	40.7	60.5	44.1	35.4	57.3	37.6
PCL (Li et al., 2021)	41.0	60.8	44.2	35.6	57.4	37.8
IFND (Ours)	**41.8**	**61.2**	**44.5**	**36.1**	**57.6**	**38.5**

- Clustering quality on ImageNet:

Method	NMI
DeepCluster (Caron et al., 2018)	43.2 ± 2.9
MoCo v2 (Chen et al., 2020d)	57.9 ± 2.2
SwAV (Caron et al., 2020)	63.8 ± 1.6
PCL (Li et al., 2021)	65.0 ± 1.9
IFND (Ours)	$\mathbf{67.5} \pm 1.7$

- Visualization of embedding space:



See More…

- Our paper:
- Code: