IMPERIAL

Unsupervised Training of Vision Transformers with Synthetic Negatives

Nikolaos Giakoumoglou, Andreas Floros, Kleanthis Marios Papadopoulos, Tania Stathaki
Department of Electrical and Electronic Engineering, Imperial College London

Introduction

This paper addresses the neglected potential of hard negative samples in self-supervised learning for vision transformers.

- Previous works explored synthetic hard negatives but not in the context of vision transformers
- We build on this observation and integrate synthetic hard negatives generated "on-the-fly" to improve vision transformer representation learning
- This simple yet effective technique notably improves the discriminative power of learned representations
- Our experiments show performance improvements for both DeiT-S and Swin-T architectures

"If intelligence is a cake, the bulk of the cake is unsupervised learning"
- Yann LeCun

Preliminaries

Contrastive Learning
Contrastive learning aims to differentiate between similar and dissimilar data pairs. Given an image x, and two augmented views, we minimize the InfoNCE loss:

$$\mathcal{L}(\mathbf{q}, \mathbf{k}, \mathcal{Q}) = -\log \frac{\exp(\mathbf{q}^\top \cdot \mathbf{k}/\tau)}{\exp(\mathbf{q}^\top \cdot \mathbf{k}/\tau) + \sum_{\mathbf{n}\in\mathcal{Q}} \exp(\mathbf{q}^\top \cdot \mathbf{n}/\tau)}$$

where $\mathbf{q}$ and $\mathbf{k}$ are the representations of two augmented views, $\mathcal{Q}$ is a set of negative samples, and τ is a temperature parameter.

Methodology

Synthetic Hard Negatives
Synthetic hard negatives provide challenging examples that help models learn more discriminative features. We generate them using an arbitrary function $\mathcal{G}$ (see SynCo [2] for details):

$$\mathbf{s} = \frac{\mathcal{G}(\mathbf{q}, \mathcal{Q}^N; \xi)}{\|\mathcal{G}(\mathbf{q}, \mathcal{Q}^N; \xi)\|_2}$$

where $\mathcal{Q}^N$ are the top N hardest negatives and ξ controls the synthesis process.

The modified InfoNCE loss is given by:

$$\mathcal{L}(\mathbf{q}, \mathbf{k}, \mathcal{Q}, \mathcal{S}) = -\log \frac{\exp(\mathbf{q}^\top \cdot \mathbf{k}/\tau)}{\exp(\mathbf{q}^\top \cdot \mathbf{k}/\tau) + Z}$$

where Z represents the negative samples:

$$Z = \sum_{\mathbf{n}\in\mathcal{Q}} \exp(\mathbf{q}^\top \cdot \mathbf{n}/\tau) + \sum_{\mathbf{s}\in\mathcal{S}} \exp(\mathbf{q}^\top \cdot \mathbf{s}/\tau)$$

The first sum represents memory-based negatives and the second sum represents synthetic negatives.

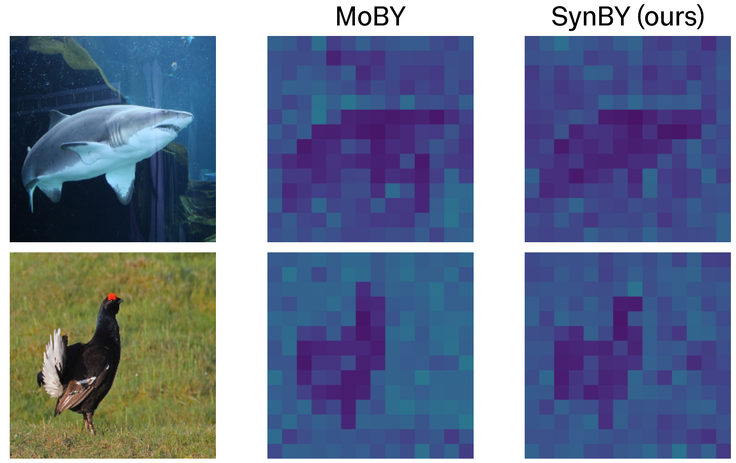
Experiments

We develop our approach in PyTorch, building upon the implementation of MoBY [1] and SynCo [2]. We pretrain on ImageNet ILSVRC-2012 using a DeiT-Small or Swin-Tiny encoder. We refer to our method as SynBY.

Linear Evaluation on ImageNet

Method	Architecture	Params (M)	Top-1 (%)
Supervised	DeiT-S	22	79.8
Supervised	Swin-T	29	81.3
DINO	DeiT-S	22	72.5
MoCo-v3	DeiT-S	22	72.5
MoBY	DeiT-S	22	72.8
MoBY	Swin-T	29	75.0
SynBY (ours)	DeiT-S	22	**73.0** ↑0.2
SynBY (ours)	Swin-T	29	**75.2** ↑0.2

Attention Visualization



References
[1] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised Learning with Swin Transformers, 2021.

[2] Nikolaos Giakoumoglou and Tania Stathaki. SynCo: Synthetic Hard Negatives for Contrastive Visual Representation Learning, 2025.