Self-Supervised Learning for Visual Representation

AIMS-DTU Research Intern Round 2

Atharv Kaushik

May 29, 2025

---

1 Introduction

This report presents the implementation and evaluation of self-supervised learning (SSL) techniques for visual representation learning. Two main paradigms were explored:

- **Contrastive Learning** using SimCLR
- **Masked Image Modeling (MIM)** using Masked Autoencoders (MAE)

Both models were pretrained on the unlabeled ImageNet-100 dataset and evaluated using linear probing.

2 Approaches Used

2.1 SimCLR

SimCLR is a contrastive learning method where different augmented views of the same image are used as positive pairs. The model is trained to bring similar (positive) examples closer and dissimilar (negative) examples apart using NT-Xent loss.

2.2 MAE (Masked Autoencoders)

MAE learns image representations by reconstructing masked patches of an input image. The encoder processes only the visible patches, while a lightweight decoder reconstructs the full image. The training objective is Mean Squared Error (MSE) on the masked pixels.

3 Model Architecture and Hyperparameters

SimCLR

- Backbone: ResNet-50
- Projection Head: 2-layer MLP

- Loss: NT-Xent (temperature = 0.5)
- Batch Size: 32
- Optimizer: Adam, Learning Rate = 1e-3
- Augmentations: Random Crop, Color Jitter, Gaussian Blur
- Epochs - 2

MAE

- Backbone: Vision Transformer (ViT-base)
- Patch Size: 16
- Masking Ratio: 75%
- Embedding Dimension: 128
- Decoder: 2-layer MLP
- Loss: Mean Squared Error (MSE)
- Batch Size: 32
- Optimizer: Adam, Learning Rate = 1e-4
- Epochs - 2

4 Evaluation Results

The performance was evaluated using a frozen encoder with a linear classifier trained on top.

Method	**Accuracy**	**F1 Score**
SimCLR	38.2%	0.36
MAE	34.7%	0.33

Table 1: Evaluation Results on ImageNet-100

5 Comparison of SimCLR and MAE

Aspect	**SimCLR**	**MAE**
Architecture	ResNet (CNN)	ViT (Transformer)
Pretraining Type	Contrastive with augmentations	Masked patch prediction
Augmentations	Required (strong)	Not required
Loss Type	NT-Xent	MSE
Training Complexity	High (needs large batch)	Lower
Representation Quality	Strong for discriminative tasks	Rich, generalized features

Table 2: Comparison of the Two SSL Approaches

6 Interpretation of Results

SimCLR showed strong performance but was sensitive to augmentations and batch size. MAE achieved competitive results with a simpler training pipeline, leveraging a transformer architecture for patch-wise prediction. MAE's flexibility makes it suitable for various vision tasks without heavy augmentation engineering.

7 References

1. Chen, Ting, et al. "A Simple Framework for Contrastive Learning of Visual Representations." *ICML 2020*. `https://arxiv.org/abs/2002.05709`

2. He, Kaiming, et al. "Masked Autoencoders Are Scalable Vision Learners." *CVPR 2022*. `https://arxiv.org/abs/2111.06377`

3. Touvron, Hugo, et al. "Training data-efficient image transformers distillation through attention." *ICML 2021*.

4. TensorFlow and PyTorch Documentation

Prepared by: Atharv Kaushik
Date: May 29, 2025