Few-Shot Representation Learning

Yating Chang, Camille Chang Electrical and Computer
Engineering University of California, Davis Davis, USA

Keywords—Self-supervised Learning, Representation Learning, Contrastive Learning, Masked Autoencoder

I. Introduction

The Oxford 102 Flowers dataset presents a challenging benchmark for image classification due to its fine-grained nature, significant intra-class variation, and extremely limited labeled training data—only ten images per class. These characteristics make it ill-suited for traditional supervised learning, which typically relies on large-scale annotated datasets to generalize well. To address these limitations, this study explores two self-supervised representation learning approaches: SimCLR and Masked Autoencoders (MAE). SimCLR is a contrastive learning framework that encourages representations of differently augmented views of the same image to be similar while pushing apart representations from different images, using the normalized temperature-scaled cross entropy (NT-Xent) loss. MAE, on the other hand, is a masked image modeling approach that reconstructs missing patches from partially observed inputs, allowing the encoder to learn globally coherent representations with minimal supervision. In our implementation, SimCLR uses a ResNet-18 backbone with a projection head trained on augmented samples, while MAE employs a Vision Transformer (ViT) encoder-decoder architecture, with 75% of image patches masked during training. Both models are implemented in PyTorch. We selected these two methods to reflect complementary learning paradigms: SimCLR is effective for learning clusterable features in the embedding space, while MAE demonstrates robustness in low-data regimes through its reconstruction-driven learning objective.

II. EXPERIMENT

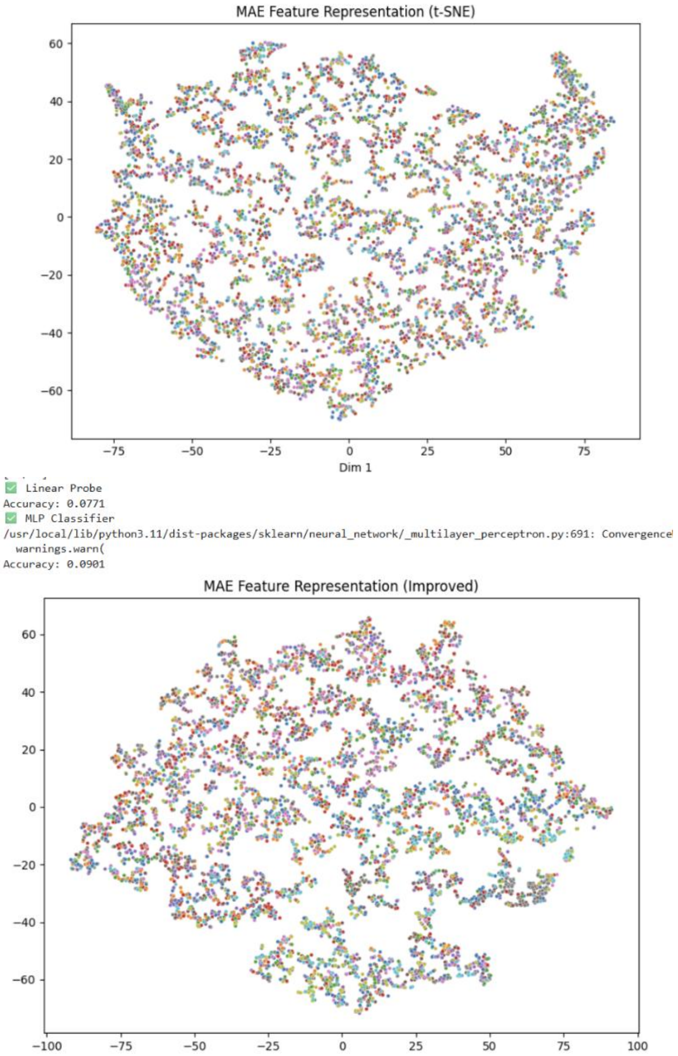
To evaluate the effectiveness of SimCLR and MAE on the Oxford 102 Flowers dataset, we designed a consistent experimental setup across both methods. All images were resized to 128×128 pixels for SimCLR and MAE, while ViT-based supervised baselines used 224×224 resolution to match model requirements. For SimCLR, we applied a sequence of strong augmentations critical for contrastive learning: random resized cropping, color jittering, Gaussian blur, random horizontal flipping, and random grayscale conversion. These augmentations were designed to produce varied views of the same image while preserving its semantic identity. MAE training did not require aggressive augmentation; instead, 75% of input patches were randomly masked during training to encourage reconstruction of missing content. SimCLR was implemented using a ResNet-18 encoder followed by a two-layer MLP projection head, trained with NT-Xent loss using the Adam optimizer (learning rate = 3e-4, batch size = 64, epochs = 20). The MAE model used a Vision Transformer encoder-decoder setup trained with MSE reconstruction loss, using Adam (learning rate = 3e-4, batch size = 64, epochs = 50). All models were trained using a single NVIDIA T4 GPU via Google Colab. To assess the learned representations, we extracted frozen features from the trained encoders and evaluated them using two downstream probes: a linear logistic regression classifier and a one-hidden-layer MLP classifier. Additionally, we visualized the feature space structure using t-SNE plots to analyze the semantic clustering capability of the learned embeddings.

III. RESULT

To evaluate the quality of learned representations, we performed linear and MLP classification probes on frozen feature embeddings extracted from SimCLR, MAE, and a supervised ViT baseline. Additionally, we visualized high-dimensional features via t-SNE.Authors and Affiliations.

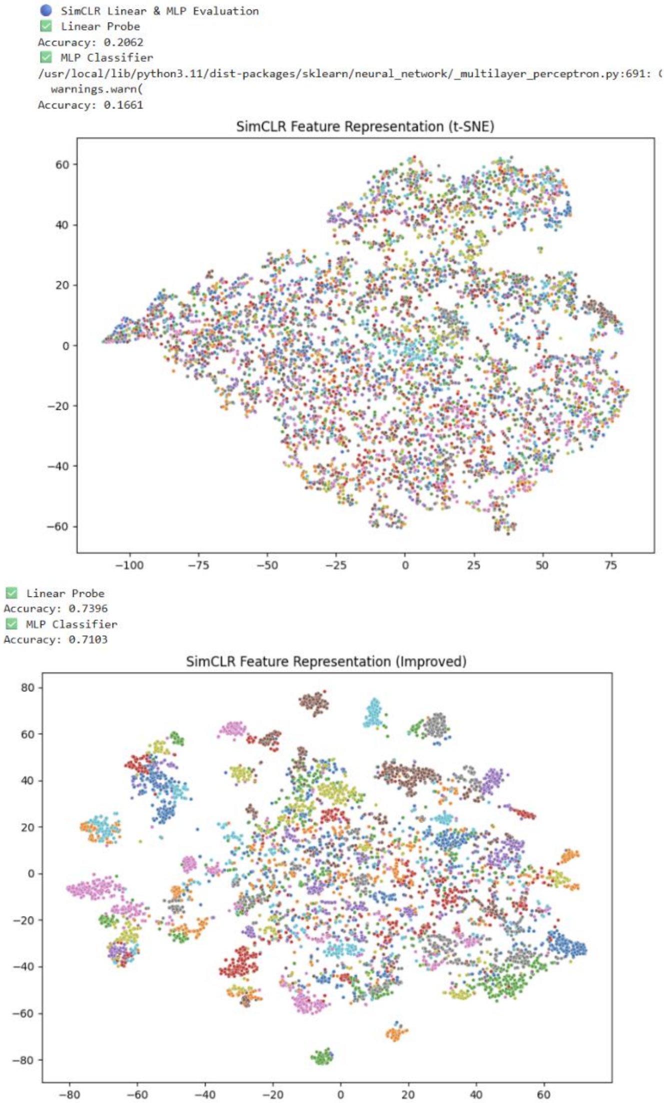
Table 1.

Method	**Linear Probe**	**MLP Classifier**
SimCLR(initial)	0.2062	0.1661
SimCLR(improved)	0.7396	0.7103
MAE(initial)	0.0790	0.0329
MAE(improved)	0.0771	0.0901
ViT (supervised)	0.9820	0.9600



```
☑ Linear Probe
Accuracy: 0.0771
☑ MLP Classifier
/usr/local/lib/python3.11/dist-packages/sklearn/neural_network/_multilayer_perceptron.py:691: Convergence
  warnings.warn(
Accuracy: 0.0901
```

Figure 1. *t-SNE visualization of MAE representations.* Up: Initial model shows scattered, poorly separated features. Down: Improved MAE with positional encoding and deeper decoder yields more structured embeddings.



```
● SimCLR Linear & MLP Evaluation
☑ Linear Probe
Accuracy: 0.2062
☑ MLP Classifier
/usr/local/lib/python3.11/dist-packages/sklearn/neural_network/_multilayer_perceptron.py:691: (
  warnings.warn(
Accuracy: 0.1661
```

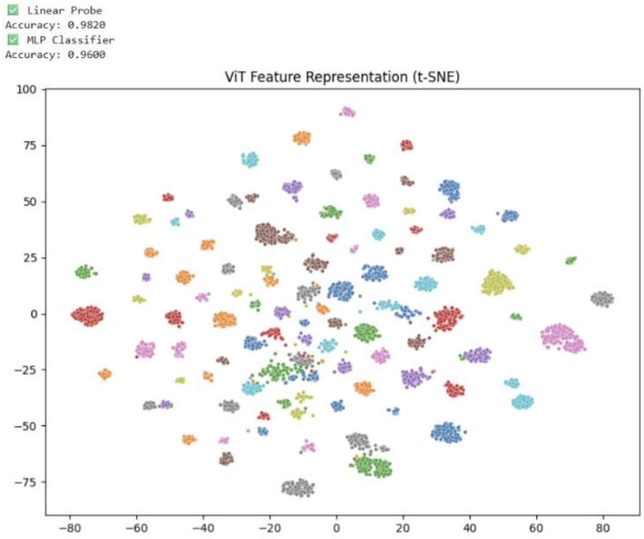
SimCLR Feature Representation (t-SNE)
60
40
20
0
−20
−40
−60
−100
−75
−50
−25
0
25
50
75

```
☑ Linear Probe
Accuracy: 0.7396
☑ MLP Classifier
Accuracy: 0.7103
```

SimCLR Feature Representation (Improved)
80
60
40
20
0
−20
−40
−60
−80
−80
−60
−40
−20
0
20
40
60

Figure 2. t-SNE visualization of SimCLR representations. (Above) Initial SimCLR model exhibits noisy, entangled clusters with poor separability. (Down) After correcting NT-Xent masking and refining augmentations, the improved SimCLR model yields well-formed and class-consistent embeddings. This aligns with the observed accuracy gains

(Linear: 20.6% → 73.9%).



```
✅ Linear Probe
Accuracy: 0.9820
✅ MLP Classifier
Accuracy: 0.9600
```

Figure 3. t-SNE visualization of ViT baseline representations. The model yields highly separable and compact clusters across classes, consistent with its strong supervised performance (Linear: 98.2%, MLP: 96.0%).

The ViT baseline, trained with full supervision, achieved near-perfect class separability, highlighting its strength in data-rich settings. In contrast, SimCLR and MAE, as self-supervised methods, demonstrated varying degrees of improvement. SimCLR showed substantial performance gains after correcting NT-Xent masking bugs and refining augmentation strategies, achieving a 74% linear probe accuracy. t-SNE plots corroborated this improvement by revealing clearer and more distinct clusters in the feature space compared to its noisy, overlapping initial version. MAE also benefitted from architectural enhancements such as the addition of decoder layers and positional encoding, leading to better-distributed and more coherent embeddings. However, despite these improvements, MAE’s clusters remained less distinct than SimCLR’s, and accuracy improvements were modest. The initial MAE representations exhibited snake-like manifolds—likely a result of reconstruction overfitting—while the improved version demonstrated more structured distributions, though still with weaker class separability. Throughout the development of our self-supervised models, several technical challenges were encountered and systematically addressed. For SimCLR, initial NT-Xent loss misalignment led to incorrect similarity computation, which we resolved by rebuilding the similarity mask to correctly exclude positive pairs. Additionally, identical contrastive views limited learning effectiveness, prompting the integration of stronger data augmentations to ensure view diversity. The MAE pipeline faced patch dimensionality mismatches between the encoder and decoder, which we corrected by refactoring the input/output shapes to match the 16×16 patchified structure. For visualization, t-SNE failed on high-dimensional inputs, so we applied spatial average pooling over patch features to obtain compact representations. The ViT baseline required resizing input images to 224×224 to match its expected dimensions. Lastly, MAE's underperformance was addressed by enhancing its decoder depth and introducing positional encodings, leading to improved spatial understanding and slightly better clustering performance.

IV. CONCLUSION

This project highlighted the strengths and limitations of self-supervised learning in low-data regimes through the lens of contrastive learning (SimCLR) and masked autoencoding (MAE). From SimCLR, we learned that contrastive loss—specifically NT-Xent—is highly effective for learning clusterable representations when diverse augmentations are applied. The method excelled in forming semantically meaningful groupings despite the absence of labels. In contrast, MAE emphasized the power of reconstructive learning, showing that even with no supervision, a model can extract spatial structure by learning to predict missing information. However, MAE struggled with class separability without additional architectural enhancements. Overall, self-supervised learning proved to be a robust strategy for representation extraction in data-scarce scenarios, bypassing the need for extensive labeling while still enabling downstream classification through simple probes like linear or MLP classifiers. These findings underscore the practical viability of self-supervised methods in real-world applications where annotation costs are high and training data is limited.

Team contrubution

Camille Chang: Focused on implementing and training the SimCLR model, including designing augmentation strategies, **tuning** contrastive loss, and analyzing clustering patterns through t-SNE. Also contributed to debugging NT-Xent loss issues and finalizing the SimCLR improvement pipeline.

Yating Chang: Led the development and evaluation of the MAE and ViT models. Responsible for architectural modifications (e.g., decoder redesign, positional encoding), MSE loss calibration, and training stability. Also handled dataset preprocessing, final report writing, and integration of experimental results.