Comparative Analysis of Latent Representations in Vision Transformer (ViT) Variants

Morteza Mahdiani, Zahra Mansouri, Yue Zhang
Machine learning 2 - Deep learning (MATH 60630A)

HEC MONTRÉAL

ABSTRACT

Vision Transformers (ViTs) leverage self-attention mechanisms to capture spatial relationships in images, often surpassing Convolutional Neural Networks (CNNs) in tasks like large-scale image classification. Despite their success in domains such as autonomous driving and medical imaging, ViTs are often criticized as "black boxes," challenging interpretability in high-stakes applications. Understanding their layer-wise latent representations is crucial for enhancing transparency and trust in real-world scenarios.

This study examines the internal activations of ViT variants—MAE, BEiT, DINOv2, AugReg, and standard ViT—on a shape-based dataset. By combining few-shot learning evaluations with Representational Similarity Matrices (RSMs) and pairwise model comparisons, we uncover how pre-training paradigms shape representational strategies. BEiT and DINOv2 demonstrate robust intermediate and deep-layer representations, with BEiT achieving the highest individual and ensemble accuracies. Multi-model ensembles further improve few-shot learning, with combinations of four and five models achieving up to 87% accuracy. These findings highlight the potential of diverse pretraining strategies and ensembles to enhance interpretability and downstream performance.

BACKGROUND

Recent advancements in representation analysis for ViTs have focused on understanding how their internal representations compare to those of convolutional neural networks (CNNs) and their implications for interpretability and performance. Raghu et al. (2022) highlighted the ability of ViTs to extract global features through self-attention, contrasting with CNNs' localized feature learning [1]. Kornblith et al. (2019) used CCA-based methods to reveal unique representational trajectories in ViTs compared to other architectures [2]. Raghu et al. (2017) introduced SVCCA to analyze feature spaces, offering insights into representational stability across transformer layers [3]. Additionally, Morcos et al. (2018) and Williams et al. (2022) proposed metrics to evaluate layer-wise similarity, demonstrating how ViTs maintain structured progressions of representations [4, 5]. These methods collectively enhance the understanding of ViTs' internal mechanisms, improving their interpretability and robustness.

Building on these insights, this study investigates how different ViT variants—such as Masked Autoencoder (MAE), BEiT, DINOv2, and AugReg—learn visual representations through diverse training strategies. While MAE uses masked image reconstruction for self-supervised learning, BEiT employs BERT-inspired masked modeling with visual tokens, and DINOv2 adopts a teacher-student self-supervised framework for robust feature learning. In contrast, standard ViTs and AugReg rely on supervised learning, with AugReg incorporating advanced data augmentation and regularization [6]. By comparing self-supervised and supervised ViT variants pre-trained on ImageNet datasets, this project examines how training strategies and fine-tuning impact latent representations and classification performance. The findings aim to improve interpretability in ViTs.

METHODOLOGY

To investigate differences in the representations of various ViT variants, we employ several analysis techniques to understand how these models process and learn features at different layers. Using the "2D Geometric Shapes Dataset" from Kaggle [1], we focus on a subset of 5 classes (circles, squares, triangles, stars, trapezoids) with 15 images per class. All models are set to evaluation mode, and input images are preprocessed to be in grayscale mode and match each model's training parameters.

Key methods include **t-SNE** for dimensionality reduction and visualization of high-dimensional latent representations across 24 transformer block layers, highlighting class separation and layer contributions [7]. **CKA (Centered Kernel Alignment)** quantifies representational similarity between layers within and across models, providing insights into how ViT variants encode input data at different depths. Representation Similarity Matrices (RSMs) are created using CKA scores [8]. We further analyze **model similarities** by computing Pearson correlations of CKA matrices across models, revealing the extent to which their internal representations align. Low correlations indicate diverse representational structures, useful for ensemble learning, while high correlations suggest interchangeable feature hierarchies [9].

Finally, a **few-shot learning** task evaluates model performance using limited training samples. Each ViT model and their combinations are fine-tuned and tested, assessing adaptability and robustness in data-constrained scenarios. These methods collectively enhance our understanding of how different ViT variants encode, represent, and generalize visual information [10].

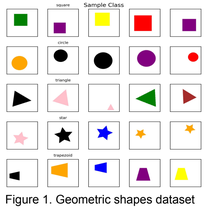

Figure 1. Geometric shapes dataset

1. https://www.kaggle.com/datasets/khalidboussaroual/2d-geometric-shapes-17-shapes

RESULTS

Representation analysis across various Vision Transformer (ViT) models reveals distinct patterns of class separation as information progresses through their layers. Using t-SNE plots, we visualize the latent representations for all models. In the **class-based t-SNE plot**, we assign distinct colors to each of the five classes (circles, squares, triangles, stars, and trapezoids) to highlight how effectively different models separate classes. For instance, DINOv2 shows strong separation by Layers 15 to 20, with compact and distinct clusters, while BEiT demonstrates significant improvement starting from Layer 8 and peaks between Layers 18 to 22.

In the **layer-based t-SNE plot**, each layer is assigned a unique color to illustrate the progression of feature learning through the 24 transformer block layers. This visualization emphasizes how early layers show minimal separation across all models, with separation improving in mid-to-late layers. For example, AugReg achieves its best separation between Layers 16 to 20, while MAE exhibits compact clusters around Layers 15 to 18, albeit less distinct than DINOv2 or AugReg. These color-coded t-SNE plots allow us to intuitively understand how each ViT variant processes and encodes class-specific features and the contributions of individual layers to the overall representation. For reference, we have shown the Tsne plot for model BEiT and DINOv2 in Figure 2 and 3.

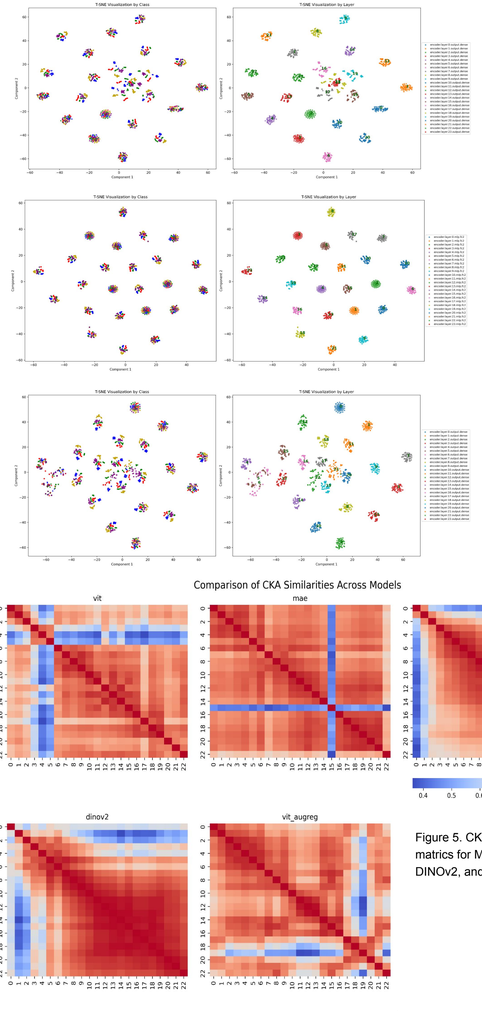

Figure 2. T-SNE visualization by classes and by layer for BEiT

Figure 3. T-SNE visualization by classes and by layer for Dinov2

Figure 4. T-SNE visualization by classes and by layer for ViT

Figure 5. CKA similarities matrics for MAE, BEiT, ViT, DINOv2, and AugReg models

Representational Similarity Matrices (RSMs) reveal distinct activation patterns for the five ViT variants when processing a shape-based dataset (triangle, circle, square, star, trapezoid). Models like MAE and BEiT exhibit block-diagonal structures in their RSMs, indicating high internal consistency within certain layer groups. DINOv2 demonstrates smooth transitions between layers, reflecting gradual feature evolution. AugReg, however, shows weaker overall similarity, likely due to its heavy augmentation and regularization strategies during training. These findings highlight diverse internal representations stemming from distinct pre-training paradigms.

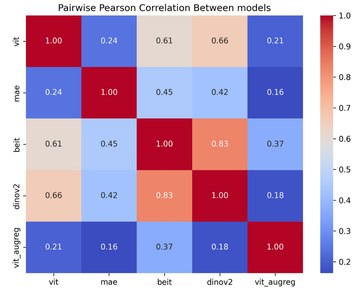

Figure 6. Pearson Correlation among CKA similarities matrics for MAE, BEiT, ViT, DINOv2, and AugReg models

Pairwise Pearson correlation analysis of activations across models demonstrates significant variation in representational overlap. BEiT and DINOv2 share the highest correlation (0.83), suggesting similar feature extraction processes driven by token-based self-supervised learning. ViT and MAE exhibit lower correlations with other models, reflecting their unique pre-training objectives (supervised for ViT, reconstruction-based for MAE). AugReg shows the weakest correlations overall, likely due to its emphasis on diversity through augmentation, leading to distinct representational characteristics.

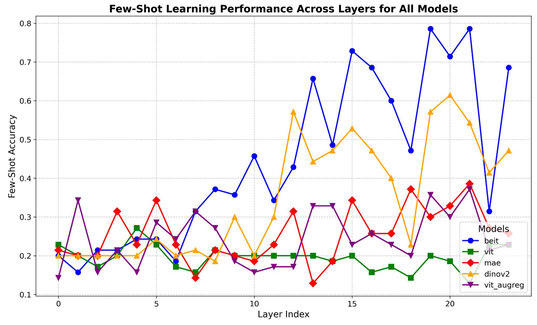

Figure 7. Layer-wise 5-shot learning accuracy for MAE, BEiT, ViT, DINOv2, and AugReg models on downstream tasks using a subset of the shape dataset.

The layer-wise 5-shot learning analysis reveals differing downstream task performances across the five ViT models. MAE and BEiT achieve peak accuracy in their deeper layers, showcasing strong feature representations in the final stages. DINOv2 performs consistently across its layers, with notable peaks in intermediate layers, indicating robust mid-layer activations. ViT maintains stable yet moderate performance across all layers, while AugReg struggles to reach comparable accuracy, likely due to its focus on diverse augmentations rather than cohesive representations.
The selected layers for each model—**BEiT** (19, 21, 15), **ViT** (4, 0, 5), **MAE** (21, 18, 5), **DINOv2** (20, 12, 19), and **AugReg** (21, 19, 1)—demonstrated the highest accuracies in the few-shot analysis. These layers were subsequently chosen for ensemble methods to leverage their strong representational capabilities and improve overall performance.

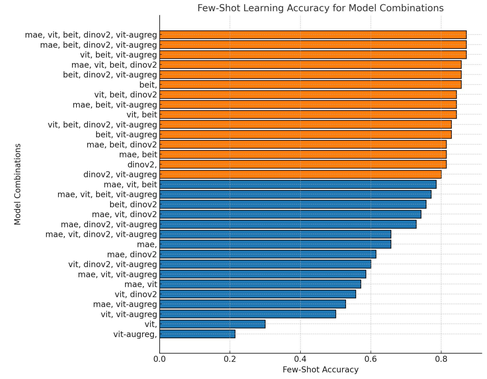

Figure 8. 5-shot learning accuracy for all possible combinations of the top three performing layers across ViT variants. Combinations include single models, pairs, triplets, as well as combinations of four and all five models. The top three layers from each model are concatenated to evaluate downstream performance.

The evaluation of model combinations, including pairs, triplets, and larger groupings, demonstrates the effectiveness of leveraging complementary feature representations. BEiT and DINOv2 consistently emerge as the most impactful contributors, achieving the highest individual and combined performance. Notably, combinations of four and five models, concatenating their top three layers, reach up to 87% accuracy, highlighting the power of representational diversity. However, combinations involving AugReg show slightly reduced performance, suggesting limited synergy with other models. These results underscore the benefits of ensembling multiple architectures for robust few-shot learning.

CONCLUSION

Our analysis reveals that different pretraining strategies significantly impact the internal representations and downstream performance of ViT models. BEiT and DINOv2 demonstrate robust feature representations and high accuracy in both individual and combined evaluations. Layer-wise analysis highlights that mid-to-deep layers are crucial for few-shot learning performance, especially in self-supervised models like BEiT and DINOv2. AugReg, while unique in its approach, exhibits weaker alignment with other models in multi-model evaluations.
The results highlight the potential of combining models with diverse training paradigms to improve downstream performance. Multi-model ensembles leveraging complementary representations achieved the highest accuracies, emphasizing the benefits of representational diversity. This finding is particularly relevant for scenarios with limited data, where feature richness from multiple sources can significantly enhance task performance.

Limitations and Future work

Our study focused on the activation patterns of Vision Transformer (ViT) variants using a shape-based dataset, limiting generalizability to texture-based or real-world images. We analyzed ViT models exclusively, excluding other architectures like Convolutional Neural Networks (CNNs) or hybrid transformers. Computational constraints restricted our experiments to a small dataset and large ViT models, potentially impacting the robustness and scalability of the results.

Future work should explore diverse datasets, including texture-based ones, and evaluate other architectures to broaden insights. Scaling experiments to larger datasets and computational resources would enhance generalizability, while investigating models of varying sizes could further refine our understanding of how architecture and pretraining influence representation learning and task performance.

References

1. M. Raghu, et al (2022)
2. S. Kornblith, et al (2019)
3. M. Raghu, et al (2017)
4. A. S. Morcos, et al (2018)
5. A. H. Williams, et al (2022)
6. Dosovitskiy, A., et al (2021)
7. He, K., et al (2022)
8. Bao, H.,et al (2022)
9. Oquab, M.,et al (2023)
10. Steiner, A., et al (2021)