AE-ViT: Token Enhancement for Vision Transformers via CNN-based Autoencoder Ensembles.

Heriniaina Andry RABOANARY[1,2], Roland RABOANARY[1,3], and Nirina Maurice HASINA TAHIRIDIMBISOA[1,4]

[1] Equipe d'accueil PNA-PHE, EDPA - Faculté des Sciences, Université d'Antananarivo, Antananarivo, Madagascar

Abstract. While Vision Transformers (ViTs) have revolutionized computer vision with their exceptional results, they struggle to balance processing speed with visual detail preservation. This tension becomes particularly evident when implementing larger patch sizes. Although larger patches reduce computational costs, they lead to significant information loss during the tokenization process. We present *AE-ViT*, a novel architecture that leverages an *ensemble of autoencoders* to address this issue by introducing *specialized latent tokens* that integrate seamlessly with standard patch tokens, enabling ViTs to capture both global and fine-grained features.

Our experiments on CIFAR-100 show that AE-ViT achieves a 23.67% relative accuracy improvement over the baseline ViT when using 16×16 patches, effectively recovering fine-grained details typically lost with larger patches. Notably, AE-ViT maintains relevant performance (60.64%) even at 32×32 patches. We further validate our method on CIFAR-10, confirming consistent benefits and adaptability across different datasets.

Ablation studies on ensemble size and integration strategy underscore the robustness of AE-ViT, while computational analysis shows that its efficiency scales favorably with increasing patch size. Overall, these findings suggest that AE-ViT provides a practical solution to the patch-size dilemma in ViTs by striking a balance between accuracy and computational cost, all within a simple, end-to-end trainable design.

Keywords: Vision Transformers, Convolutional Neural Networks, Autoencoders, Hybrid Architecture, Image Classification, latent representation

1 Introduction

Vision Transformers (ViTs) [9] have emerged as a powerful alternative to Convolutional Neural Networks (CNNs) in computer vision tasks. Adapting the transformer architecture from natural language processing [26], Vision Transformers (ViTs) process images by decomposing them into non-overlapping patches and employing *self-attention* mechanisms to model relationships across the entire image. This architectural paradigm has achieved exceptional performance across diverse computer vision tasks [14], presenting a compelling challenge to the historically dominant Convolutional Neural Networks (CNNs).

However, ViTs face a critical trade-off in patch size selection. The computational complexity of self-attention operations grows quadratically with the number of tokens, making smaller patches (e.g., 8×8 pixels) computationally expensive despite their rich feature representation. Conversely, larger patches (e.g., 16×16 pixels) offer better computational efficiency but sacrifice fine-grained spatial information [19]. This information loss becomes particularly evident in tasks requiring detailed feature analysis [4], where the tokenization process with large patches might lose crucial local patterns.

Several approaches have been proposed to address this trade-off. Hierarchical designs [27] progressively merge tokens to balance computational cost and feature granularity. Efficient attention mechanisms [20] reduce complexity through local attention windows.

DOI:10.5121/ijaia.2025.16104

Hybrid architectures [10] combine CNNs and transformers to leverage their complementary strengths. However, these solutions often introduce significant architectural complexity [12], requiring careful design choices and sophisticated training strategies that may limit their practical adoption.

In this paper, we introduce AE-ViT, a novel approach that addresses the patch size dilemma through an ensemble of autoencoders. Our method complements the standard patch-based tokenization with learned latent representations from convolutional autoencoders [1]. These autoencoders, operating at a finer scale than the patch tokens, capture and preserve local features that would otherwise be lost with large patches. By integrating these latent representations into the transformer's token sequence, we enable the model to simultaneously leverage both the computational benefits of large patches and the fine-grained feature detection capabilities of CNNs [18]. This hybrid architecture builds upon the proven strengths of both convolutional [17] and transformer architectures [24], creating a synergy that effectively compensates for the limitations of large patch tokenization.

Our key contributions can be summarized as follows: (1) We propose a novel hybrid architecture that leverages an ensemble of autoencoders to compensate for information loss in large-patch Vision Transformers; (2) We introduce an efficient method for integrating autoencoder latent representations with transformer tokens, maintaining architectural simplicity while significantly improving performance; (3) We demonstrate the effectiveness of our approach through extensive experiments on CIFAR-100 [15], achieving a 23.67% relative accuracy improvement over the baseline ViT when using 16×16 patches, without introducing significant computational overhead [21]. We also tested our system on the CIFAR-10 [16] to validate the scaling over other datasets. Our results show that AE-ViT provides a practical solution to the patch size dilemma, particularly valuable in scenarios where computational efficiency is crucial [9].

2 Background and Related Work

2.1 Vision Transformers

Originally introduced by Dosovitskiy et al. [9], Vision Transformers (ViTs) adapt the transformer architecture [26] for image processing tasks by treating images as sequences of patches. This approach employs self-attention mechanisms to capture global dependencies, yielding remarkable performance across numerous vision applications. The success of the original ViT has led to multiple improvements, such as DeiT [24], which proposes efficient training techniques, and CaiT [25], which enhances feature representation by increasing model depth.

2.2 Efficiency in Vision Transformers

Despite their success, ViTs face computational challenges stemming from self-attention complexity, which grows quadratically with the number of tokens. Smaller patches yield more tokens and capture finer details, but with significantly higher computational cost. To address this, the Swin Transformer [20] introduces local attention windows to reduce overall complexity, and PVT [27] employs a progressive shrinking pyramid to reduce the token count in deeper layers. Another approach, TNT [13], processes patch-level and pixel-level tokens in a nested structure, emphasizing multi-scale feature representation.

2.3 Hybrid CNN–Transformer Approaches

Beyond pure transformer architectures, a growing body of research integrates convolutional neural networks (CNNs) into ViTs to capitalize on their complementary strengths. ConViT [7] introduces soft convolutional inductive biases to refine local feature modeling, while LeViT [10] interleaves convolutional blocks for efficient early-stage processing. CvT [28] demonstrates the effectiveness of convolutional token embeddings, and early convolution layers [29] have been shown to be crucial for robust vision transformer performance. Meanwhile, CoAtNet [6] unifies depthwise convolutions and attention layers for scalable performance across different input sizes, and MobileViT [22] incorporates lightweight CNN blocks into ViTs to facilitate mobile-friendly inference. In parallel, Guo et al. [11] highlight that CNN-based approaches can significantly reduce the computational costs of vision transformers. These hybrid efforts underscore the synergy between local receptive fields (CNN-like) and global self-attention (transformer-like), paving the way for more efficient and effective models.

2.4 Autoencoders in Vision Tasks

Autoencoder architectures [1] have shown remarkable success in learning compact, meaningful representations, from dimensionality reduction to feature learning [2]. Their ability to preserve essential information while reducing spatial dimensionality is particularly useful in tasks requiring fine-grained detail [8]. By reconstructing input data, autoencoders can capture local and global structure, making them relevant for scenarios where large image patches risk losing crucial spatial information.

2.5 Our Approach

Our work bridges these lines of research by introducing an *ensemble* of CNN-based autoencoders to enhance transformer-based vision models. While previous methods have incorporated CNN blocks within a transformer (Section 2.3) or tackled patch-level efficiency (Section 2.2), our method uniquely *leverages autoencoder-based latent representations* to compensate for information loss in large-patch ViTs. Instead of modifying the patch embedding or internal attention blocks, we integrate a learnable latent "AE token" that complements the standard patch tokens. This design aims to reconcile the efficiency benefits of large patches with the need to preserve fine-grained details, offering a novel solution to the efficiency–accuracy trade-off in modern vision transformers.

3 Method

3.1 Overview

The Fig. 1 presents an overview of the architecture of our ensemble. For clarity, we chose to present the case of 4 autoencoders. The latent token is "*plugged*" as an additional token for the transformer.

3.2 Autoencoder Ensemble Design

Each autoencoder in our ensemble follows a convolutional architecture optimized for capturing fine-grained features. The encoder pathway consists of three downsampling blocks that progressively reduce spatial dimensions from 64×64 to 8×8 while increasing the feature channels. Specifically, we use strided convolutions with a kernel size of 4×4 and stride

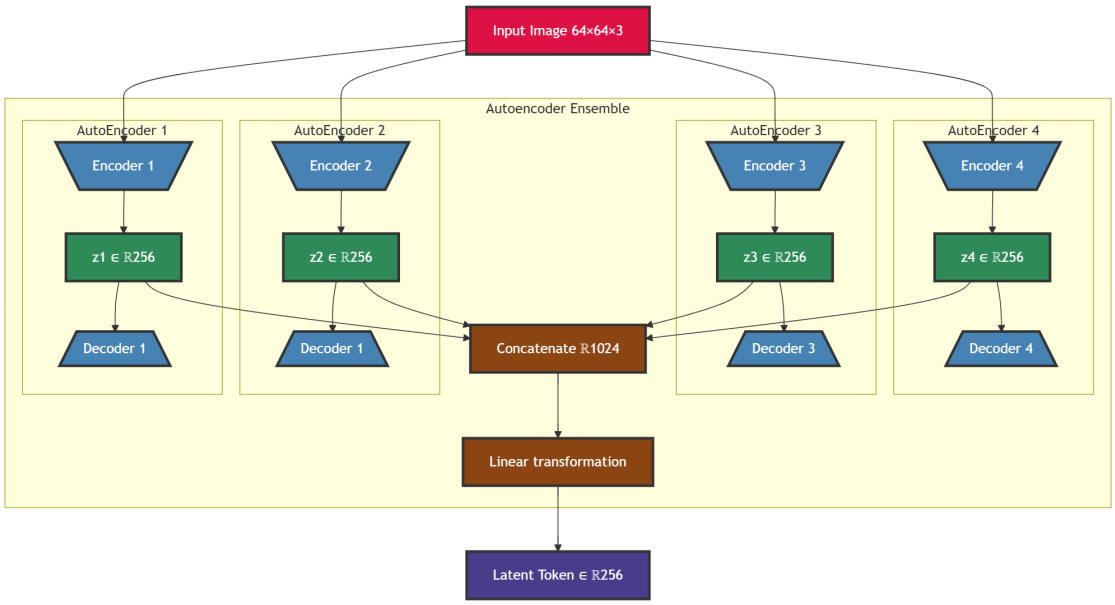

Fig. 1. Overview of the ensemble architecture (for four autoencoders).

2, followed by batch normalization and ReLU activation. The feature dimensions evolve as follows:

$$3 \xrightarrow{\text{conv}} 32 \xrightarrow{\text{conv}} 64 \xrightarrow{\text{conv}} 128 \tag{1}$$

The final feature map is flattened and projected to a latent space of dimension 256 through a fully connected layer. This results in a compression ratio of:

$$\text{compression ratio} = \frac{64 \times 64 \times 3}{256} = 48 : 1 \tag{2}$$

The decoder mirrors this structure with transposed convolutions to progressively reconstruct the spatial dimensions:

$$128 \xrightarrow{\text{conv}^\mathsf{T}} 64 \xrightarrow{\text{conv}^\mathsf{T}} 32 \xrightarrow{\text{conv}^\mathsf{T}} 3 \tag{3}$$

To prevent overfitting and encourage robust feature learning, we employ L1 and L2 regularization on the latent space:

$$\mathcal{L}_{\text{reg}} = \lambda_1 \|\mathbf{z}\|_1 + \lambda_2 \|\mathbf{z}\|_2^2 \tag{4}$$

where $\mathbf{z}$ represents the latent vector and $\lambda_1 = \lambda_2 = 0.01$ are regularization coefficients. The total loss for each autoencoder is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{reg}} \tag{5}$$

where $\mathcal{L}_{\text{recon}}$ is the mean squared error between the input and reconstructed images.

3.3 Latent Token Integration

The main innovation of our approach lies in how we integrate the autoencoder ensemble's latent representations with the Vision Transformer's patch tokens. Given an input image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$, the process follows three parallel paths that merge in the transformer:

Patch Tokenization The input image is divided into non-overlapping patches of size $s_p \times s_p$, resulting in $N = (H/s_p) \times (W/s_p)$ patches. Where s_p defines the patch size in pixels. Each patch is linearly projected to dimension D through a learnable embedding matrix $\mathbf{E} \in \mathbb{R}^{(s_p \times s_p \times 3) \times D}$:

$$\mathbf{t}_{\text{patch}} = [\mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; ...; \mathbf{x}_p^N\mathbf{E}] \in \mathbb{R}^{N \times D} \tag{6}$$

Latent Encoding Simultaneously, our ensemble of five autoencoders processes the input image, each producing a latent vector $\mathbf{z}_i \in \mathbb{R}^{256}$. These latents are concatenated:

$$\mathbf{z}_{\text{concat}} = [\mathbf{z}_1; \mathbf{z}_2; \mathbf{z}_3; \mathbf{z}_4; \mathbf{z}_5] \in \mathbb{R}^{1280} \tag{7}$$

This concatenated representation is then projected to the transformer's embedding dimension D through a learnable projection $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{1280 \times D}$:

$$\mathbf{t}_{\text{latent}} = \mathbf{z}_{\text{concat}}\mathbf{W}_{\text{proj}} \in \mathbb{R}^{1 \times D} \tag{8}$$

Token Sequence Formation The final sequence presented to the transformer concatenates the class token $\mathbf{t}_{\text{cls}}$, patch tokens, and latent token:

$$\mathbf{T} = [\mathbf{t}_{\text{cls}}; \mathbf{t}_{\text{patch}}; \mathbf{t}_{\text{latent}}] \in \mathbb{R}^{(N+2) \times D} \tag{9}$$

Position embeddings $\mathbf{P}$ are added to this sequence to maintain positional information:

$$\mathbf{T}_{\text{final}} = \mathbf{T} + \mathbf{P} \tag{10}$$

For the final classification, we utilize both the class token and the latent token, concatenating their representations after the transformer processing:

$$\mathbf{y} = \text{MLP}([\mathbf{h}_{\text{cls}}; \mathbf{h}_{\text{latent}}]) \tag{11}$$

where $\mathbf{h}_{\text{cls}}$ and $\mathbf{h}_{\text{latent}}$ are the transformed representations of the respective tokens.

3.4 Training Strategy

Our training process follows a two-phase approach designed to maximize the complementary strengths of both the autoencoder ensemble and the Vision Transformer.

Phase 1: Autoencoder Ensemble Pretraining We train each autoencoder independently on different random subsets of the training data. Each autoencoder sees different parts of the training data, randomly sampled with a fixed seed to ensure reproducibility. The objective function for each autoencoder combines reconstruction loss with latent space regularization:

$$\mathcal{L}_{\text{AE}} = \mathcal{L}_{\text{recon}} + \lambda_1\|\mathbf{z}\|_1 + \lambda_2\|\mathbf{z}\|_2^2 \tag{12}$$

where $\lambda_1 = \lambda_2 = 0.01$. This phase runs for 5 to 20[5] epochs using the Adam optimizer with a learning rate of 10^{-3}.

[5] Depending on the experience conditions.

Phase 2: End-to-End Training After pretraining, we freeze the autoencoder parameters and train the complete AE-ViT architecture end-to-end. The loss function for this phase is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{wd}} \|\theta\|_2^2 \tag{13}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss with label smoothing ($\alpha = 0.1$), and $\lambda_{\text{wd}} = 0.05$ is the weight decay coefficient. We employ the AdamW optimizer with the following specifics:

- Learning rate: 10^{-3}
- Batch size: 64
- Training epochs: 100
- No early stopping, use the last epoch for test
- Cosine learning rate scheduling
- Gradient clipping at norm 1.0

The complete training process can be formalized as:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}_{\text{total}}(\text{AE-ViT}_\theta(x), y)] \tag{14}$$

where $\mathcal{D}$ represents the training dataset and θ are the model parameters.

4 Experiments

4.1 Experimental Setup

We conduct extensive experiments on two standard benchmarks: CIFAR-100 [15] and CIFAR-10 [16]. Images are resized to 64×64 pixels. We use standard data augmentation including random horizontal flips and AutoAugment. Training uses AdamW optimizer with learning rate 1e-3 and weight decay 0.05 for 100 epochs with cosine scheduling.

We conduct our experiments on CIFAR-100, which consists of 50,000 training images and 10,000 test images in 100 classes. All images are resized to 64×64 pixels to accommodate our patch-based architecture. We use standard data augmentation techniques including random horizontal flips and AutoAugment [5]. Implementation details are provided in our publicly available code.

4.2 Main Results

Performance on CIFAR-100. Table 1 presents our main results with various patch sizes:

Table 1. Classification accuracy (%) on CIFAR-100

Model	Patch Size	FLOPs (M)	Accuracy
ViT	8×8	287.6	61.23
ViT	16×16	78.4	49.94
ViT	32×32	19.6	35.86
AE-ViT (Ours)	16×16	187.0	61.76
AE-ViT (Ours)	32×32	128.4	60.64

With 16×16 patches, our approach not only outperforms the baseline ViT but also achieves better accuracy than the more computationally intensive 8×8 patch configuration,

and has 23.67% relative accuracy improvement compared to baseline ViT with 16x16 patches. Most notably, with 32×32 patches, AE-ViT maintains reasonable performance (60.64%) while the baseline ViT severely degrades (35.86%), demonstrating a remarkable relative improvement of 69%.

Scaling to CIFAR-10. To validate the generality of our approach, we conduct some experiments on CIFAR-10. Results in Table 2 show that AE-ViT maintains its effectiveness:

Table 2. Classification accuracy (%) on CIFAR-10

Model	Patch Size	Accuracy
ViT	8×8	82.91
ViT	16×16	76.37
AE-ViT (Ours)	16×16	**84.35**

We can see that the architecture also works on the CIFAR-10 dataset.

4.3 Ablation Studies

Impact of Ensemble Size. We conduct a systematic study of ensemble size impact:

Table 3. Impact of number of autoencoders (CIFAR-100)

#AEs	Test Acc.	Train Acc.	FLOPs (M)
1	59.11	86.28	106
2	60.73	87.15	133
3	61.00	88.38	160
4	**61.76**	88.90	187
5	61.75	89.31	214
6	61.27	89.08	241

The results reveal an optimal *configuration* at 4 autoencoders, beyond which performance plateaus or slightly degrades. This suggests that while ensemble diversity is beneficial, there exists a sweet spot balancing performance and computational cost. However, we should conduct more experiments on various conditions (dataset, patch size, transformers hyperparameters) before we can *generalize* this *configuration.*

Cross-dataset Training. We explore enhancing the ensemble with an autoencoder trained on CIFAR-10. This configuration achieves 61.94% accuracy on CIFAR-100, suggesting potential benefits from cross-dataset knowledge transfer, albeit with diminishing returns compared to the computational overhead.

Conclusion. These results indicate that adding more than 4 autoencoders is not beneficial for the AE-ViT architecture. The use of 4 autoencoders represents an optimal balance between accuracy, generalization, and computational efficiency.

4.4 Efficiency Analysis

While our primary results demonstrate AE-ViT's superior accuracy over the baseline ViT with 16×16 patches, a more insightful comparison emerges when we consider the baseline

ViT with 8×8 patches, which achieves similar performance levels. This comparison is particularly interesting as it addresses the trade-off between accuracy and computational efficiency.

Table 4 presents a detailed comparison:

Table 4. Efficiency comparison between ViT (8×8) and AE-ViT (16×16)

Model	FLOPs (M)	#Tokens	Test Acc. (%)
ViT (8×8)	287.6	65	61.23
AE-ViT (16×16)	187.0	18	61.76
Relative Difference	-34.9%	-72.3%	+0.87%

4.5 Comparison with State-of-the-Art

While our primary focus is on addressing the patch size dilemma in Vision Transformers, it is insightful to position AE-ViT within the broader context of modern architectures. Table 5 presents a comparison with Swin Transformer [20], a leading hierarchical vision transformer:

Table 5. Comparison with Swin Transformer on CIFAR-100

Model	Accuracy (%)	Params (M)	FLOPs (M)
Swin-T [3]	78.41	28.3	4500
AE-ViT (4 AEs)	61.76	12.4	187

While Swin Transformer achieves higher accuracy, AE-ViT offers significantly better computational efficiency.

This efficiency gap becomes even more significant when considering higher resolution images. For example, with 1080p images (1920×1080):

Table 6. Theoretical scaling to 1080p images

Model	Memory	FLOPs (G)
Swin-T	$O(HW)$	284.4
AE-ViT	$O((HW/P^2))$	11.8

This scaling advantage stems from our efficient use of large patches (16×16) combined with the fixed-cost autoencoder ensemble. While Swin Transformer's computational requirements grow quadratically with image size, AE-ViT maintains better efficiency, making it particularly suitable for high-resolution applications where computational resources are constrained.

The results reveal that AE-ViT achieves comparable (slightly better) accuracy while requiring about 35% fewer FLOPs. This efficiency gain stems primarily from two factors:

1) Token Efficiency: AE-ViT processes only 18 tokens (16 patch tokens + 1 CLS token + 1 latent token) compared to 65 tokens in the 8×8 ViT, resulting in a 72.3% reduction in the self-attention computational load.

2) Computational Distribution: While AE-ViT introduces additional computation through its autoencoder ensemble (136.3M FLOPs), this is more than offset by the reduced transformer complexity (78.4M FLOPs vs 287.6M FLOPs).

The memory footprint also favors AE-ViT, as the attention mechanism's quadratic memory scaling with respect to the number of tokens ($O(n^2)$) makes the reduced token count particularly significant. This demonstrates that our approach not only bridges the performance gap of larger patches but does so in a computationally efficient manner.

5 Discussion

Our experimental results demonstrate several key findings about AE-ViT and provide insights into the trade-offs between patch size, computational efficiency, and model performance.

5.1 Optimal Ensemble Configuration

The systematic study of autoencoder ensemble size reveals a clear optimization pattern. Starting from a single autoencoder (59.11%), we observe significant improvements with each additional autoencoder up to four (61.76%), followed by diminishing returns with five autoencoders (61.75%) and performance degradation with six (61.27%). This pattern suggests that:

- The ensemble approach is fundamentally sound, with even two autoencoders outperforming a single autoencoder by 1.62%
- Four autoencoders represent an optimal balance between performance and complexity
- Additional autoencoders beyond four may introduce unnecessary redundancy or noise

5.2 Scaling with Patch Size

Perhaps the most striking result is AE-ViT's ability to maintain performance with larger patch sizes:

- With 16×16 patches, AE-ViT (61.76%) matches the performance of standard ViT with 8×8 patches (61.23%) while using 35% fewer FLOPs
- With 32×32 patches, AE-ViT (60.64%) demonstrates remarkable resilience compared to the baseline (35.86%), achieving a 69% relative improvement

This scaling behavior suggests that our autoencoder ensemble effectively compensates for the information loss in larger patches, potentially offering a pathway to processing high-resolution images efficiently.

5.3 Cross-Dataset Insights

Our experiments with cross-dataset training, where we incorporate an autoencoder trained on CIFAR-10 into the ensemble, yield several insights:

- The mixed ensemble (61.94%) slightly outperforms the pure CIFAR-100 ensemble (61.76%)
- The improvement, while modest, suggests potential benefits from diverse training data
- The approach maintains effectiveness across datasets, as demonstrated by strong performance on CIFAR-10 (84.35%)

5.4 Computational Efficiency

Comparing AE-ViT with state-of-the-art models like Swin Transformer reveals an interesting efficiency-accuracy trade-off:

- While Swin-T achieves higher accuracy (78.41%), it requires 24× more FLOPs
- AE-ViT's efficiency advantage grows with image resolution due to fixed autoencoder costs
- The parameter count remains modest (12.4M vs 28.3M for Swin-T)

5.5 Limitations

Our approach, while effective, has several notable limitations:

- Current architecture depends on a fixed number of autoencoders, lacking adaptability to varying computational constraints
- Performance on high-resolution images remains untested, particularly for resolutions beyond 64×64
- The approach has only been validated on classification tasks, not on more complex vision tasks
- Limited to static image processing, without consideration for temporal features
- Generalization capabilities have only been tested on relatively small datasets (CIFAR-100, CIFAR-10)

5.6 Future Work

Several promising directions for future research emerge:

Architecture Improvements

- Develop dynamic ensemble selection mechanisms that adapt to specific domains and computational constraints
- Explore alternative autoencoder architectures for enhanced feature extraction
- Investigate integration with state-of-the-art transformer variants
- Explore cross dataset learning seen on 4.3 for further knowledge transfer.
- Using various types of autoencoders to increase the robustness of transformers. To have better results than those seen in [23]

Scaling and Performance

- Validate performance on high-resolution images (1080p, 4K)
- Test scalability with larger datasets (ImageNet, Places365)
- Optimize implementation for specific hardware accelerators (GPUs, TPUs)

Extended Applications

- Adapt the architecture for dense prediction tasks (segmentation, detection)
- Extend to video processing by incorporating temporal information
- Explore transfer learning for specialized domains (medical imaging, satellite imagery)

6 Conclusion

In this paper, we introduced AE-ViT, a novel approach that effectively addresses the patch-size dilemma in Vision Transformers through an ensemble of autoencoders. Our method achieves a 23.67% relative improvement over the baseline ViT on CIFAR-100 with 16×16 patches, while using significantly fewer computational resources than models that rely on smaller patches. Through extensive experimentation, we identified an optimal configuration of four autoencoders that balances performance and efficiency.

The effectiveness of AE-ViT is particularly evident in its ability to maintain strong performance with large patches, achieving 60.64% accuracy with 32×32 patches compared to the baseline's 35.86%. This capability, combined with its strong showing on CIFAR-10 (84.35%) and efficient scaling properties, demonstrates the potential of our approach for high-resolution image processing applications. Most importantly, AE-ViT offers a practical solution for scenarios requiring efficient vision processing, providing a favorable trade-off between accuracy and computational cost. While more computationally intensive models like Swin Transformer achieve higher absolute accuracy, AE-ViT's efficiency (24× fewer FLOPs) makes it particularly attractive for resource-constrained settings or real-time processing of high-resolution images.

In line with our findings, promising avenues for further research include exploring dynamic ensemble selection mechanisms to adapt AE-ViT to specific domains, developing more sophisticated autoencoder architectures for enhanced feature extraction, and scaling to larger datasets like ImageNet or higher resolutions (1080p and beyond). Extending AE-ViT to dense prediction tasks (e.g., segmentation or detection) and time-series data (e.g., video) would also be valuable, as would investigating transfer learning for specialized domains such as medical imaging or satellite imagery. These extensions may reinforce AE-ViT's robustness, accelerate its adoption in real-world scenarios, and deepen our understanding of hybrid autoencoder–transformer models.

7 Acknowledgments

We thank the anonymous reviewers for their valuable feedback. We thank as well ISPM (`https://ispm-edu.com/`) for funding our research.

References

1. Baldi, P.: Autoencoders, unsupervised learning, and deep architectures. Proceedings of ICML Workshop on Unsupervised and Transfer Learning (2012)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(8), 1798–1828 (2013)
3. Chen, C., Derakhshani, M.M., Liu, Z., Fu, J., Shi, Q., Xu, X., Yuan, L.: On the vision transformer scaling: Parameter scaling laws and improved training strategies. arXiv preprint arXiv:2204.08476 (2022)
4. Chen, C.F., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
5. Cubuk, E.D., Zoph, B., Mané, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 113–123 (2019). https://doi.org/10.1109/CVPR.2019.00020
6. Dai, Z., Liu, H., Le, Q.V., Tan, M.: CoAtNet: Marrying Convolution and Attention for All Data Sizes. In: Advances in Neural Information Processing Systems. vol. 34, pp. 3965–3977 (2021)
7. d'Ascoli, S., Touvron, H., Leavitt, M., Morcos, A., Biroli, G., Sagun, L.: Convit: Improving vision transformers with soft convolutional inductive biases. In: Proceedings of the International Conference on Machine Learning (ICML) (2021)

8. Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B.: Attention in attention: Modeling context correlation for efficient vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
10. Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., Douze, M.: Levit: A vision transformer in convnet's clothing for faster inference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
11. Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., Xu, C.: Cmt: Convolutional neural networks meet vision transformers (2022), https://arxiv.org/abs/2107.06263
12. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al.: A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
13. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y.: Transformer in transformer. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
14. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM Computing Surveys (2021)
15. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009)
16. Krizhevsky, A., Nair, V., Hinton, G.: CIFAR-10 (canadian institute for advanced research). http://www.cs.toronto.edu/~kriz/cifar.html (2009), dataset
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems **25** (2012)
18. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
19. Liu, K., Zhang, W., Tang, K., Li, Y., Cheng, J., Liu, Q.: A survey of vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
20. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Efficient transformers: A survey. ACM Computing Surveys (2021)
22. Mehta, S., Rastegari, M.: MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. In: International Conference on Learning Representations (ICLR) (2022), https://openreview.net/forum?id=7RkGYOFtwrY
23. Raboanary, H.A., Raboanary, R., Tahiridimbisoa, N.H.M.: Robustness assessment of neural network architectures to geometric transformations: A comparative study with data augmentation. In: 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME). IEEE, Tenerife, Canary Islands, Spain (Jul 2023). https://doi.org/10.1109/ICECCME57830.2023.10253075
24. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: Proceedings of the International Conference on Machine Learning (ICML) (2021)
25. Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H.: Going deeper with image transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
27. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
28. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
29. Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., Girshick, R.: Early convolutions help transformers see better. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)