ViT Registers and Fractal ViT

Anonymous authors
Paper under double-blind review

Abstract

Drawing inspiration from recent findings including surprisingly decent performance of transformers without positional encoding (NoPE) in the domain of language models and how registers (additional throwaway tokens not tied to input) may improve the performance of large vision transformers (ViTs), we invent and test a variant of ViT called fractal ViT that breaks permutation invariance among the tokens by applying an attention mask between the regular tokens and "summary tokens" similar to registers, in isolation or in combination with various positional encodings. These models do not improve upon the baseline performance, highlighting the fact that these findings may be scale, domain, or application-specific.

1 Introduction

Vision Transformer (ViT, Dosovitskiy (2020)) has emerged as a strong alternative to CNNs (convolutional neural networks) for computer vision tasks. Based on the transformer (Vaswani, 2017) architecture, it is nearly identical to transformer-based language models (LMs), except that the input tokens are linear projections of pixel patches instead of token embeddings. Similar to encoder LMs, ViT needs to break the permutation invariance of tokens with positional encoding, which has now gone through countless iterations.

While it has always been known that generative LMs based on transformer *decoder*, transformer with causal mask, do not exhibit permutation invariance, it was only reported recently that LMs based on transformer decoder without any positional encoding perform surprisingly well. Known as NoPos (Haviv et al., 2022) or NoPE, it was later shown that in the limit of infinite precision positional info can be fully reconstructed with causal mask as an explanation for its performance (Kazemnejad et al., 2023). We therefore wonder whether similar mask-based positional encoding is possible for ViT. However, preliminary experiments show that applying attention mask to regular tokens destroys performance.

Finally, it was shown recently that a small portion of tokens with a very high norm (outlier tokens) emerge in large ViT after training, which can be mitigated with the addition of throwaway tokens called registers (Darcet et al., 2024) not tied to the input or contributing to the output. This finding inspires us to test whether we use similar tokens not tied to the input and apply attention mask to them to provide positional info, without changing the all-pair attention of regular tokens.

2 Background and Related Work

2.1 Positional Encoding

There has been many variants of positional encoding for ViT. The original ViT uses learned positional encoding (Dosovitskiy, 2020), which may have contributed to its popularity among models including OpenCLIP (Ilharco et al., 2021), DEIT-III (Touvron et al., 2022), and DINOv2 (Oquab et al., 2024). Experiments reported in Darcet et al. (2024) that add registers to these three models follow the same practice and use randomly initialized, learned positional encoding for the registers (@TimDarcet, 2023). Chen et al. (2021b) proposes a 2D variant of sinusoidal positional encoding of Vaswani (2017), sincos2d, which is found to improve the performance of the ImageNet-1k ViT-S/16 baseline (Beyer et al., 2022). More recently, other positional encodings from the domain of LMs have been ported and tested in ViT, including ALiBi (Attention with Linear Biases, Press et al.

(2021)) and RoPE (Rotary Positional Embeddings, Su et al. (2021)), which give rise to 2D-ALiBi used in CROMA (Fuller et al., 2024) and RoPE-ViT (Heo et al., 2025), respectively.

2.2 Attention Patterns

While designed for the purpose of positional encoding, ALiBI can also be considered and implemented as a soft attention mask that reduces attention scores of distant query-key pairs. Conversely, while the causal mask for transformer decoder is originally designed to preserve causality of the output (Vaswani, 2017), it is shown later that decoder LMs perform surprisingly well with just the causal mask and without further positional encoding (Haviv et al., 2022; Kazemnejad et al., 2023). Other modifications of the baseline all-pair attention pattern are usually for the purpose of representation learning, improvement of the training or inference dynamics, or more compute-efficient attention mechanism:

1. Representation learning: The practice of adding a special [CLS] token for representation learning goes back to BERT (Bidirectional Encoder Representations from Transformers, Devlin et al. (2019)). This is followed by the original ViT (Dosovitskiy, 2020) and vision-language models such as CLIP (Radford et al., 2021).
2. Improvement of dynamics: Adding throwaway registers to large ViT to eliminate the artifacts of high-norm outlier tokens (Darcet et al., 2024) falls into this category. The discovery of ”attention sink”: unusually high attention score of the initial token with no semantic relevance and the mitigation of always keeping the initial token key and value for the sliding window of StreamingLLM (Xiao et al., 2024) can be considered distant parallel in the domain of LMs.
3. Compute-efficiency: Finally, efficiency of the transformer can be improved through sparsity of the attention mechanism, either with fixed attention pattern or context-dependent token dropping. People have performed extensive experiments with ViT with a variety of attention sparsity (Chen et al., 2021a; 2023), but to our best knowledge the sparse attention pattern has always been applied to the input tokens only instead of additional tokens such as registers.

3 Methodology

In fractal ViT, we add “summary tokens” not tied to the input just like registers, in addition to global token [CLS]. We then assign $k \times k$ regular tokens to each summary token (“k-summary” in our terminology) depending on the location of the tokens and apply an attention mask to break permutation invariance with a self-similar pattern:

1. All $n \times n$ regular tokens attend each other.
2. All $\frac{n}{k} \times \frac{n}{k}$ summary tokens attend each other, but each summary tokens only attends the $k \times k$ regular tokens assigned to it. The assigned regular tokens also attend back to their shared summary token.
3. Finally, the global token still attends to all tokens and all tokens still attend back to the global token.

The experiments presented in this paper exclusively focus on the simplest case where we have $k^2 \times k^2$ regular tokens, $k \times k$ summary tokens, and one global token, but this pattern can be extended to multiple levels of summary tokens. See Figure 1 for a diagram and Appendix A for an implementation.

4 Results

Here we report top-1 validation set accuracy after training a ViT-S/16 for 90 epochs on ImageNet-1k with input resolution 256 (Table 1). Except for the input resolution, positional encoding, and the addition of register or summary/global tokens, we follow the setup of (Beyer et al., 2022). The “variant” column consists of the positional encoding (sincos2d, learned, 2D-ALiBi, or none: see Appendix B), additional tokens (17 registers with learned positional encoding as in (Darcet et al.,

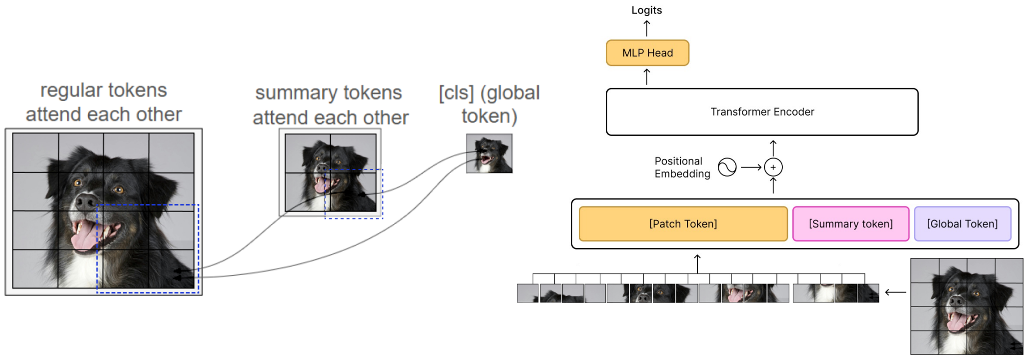

Figure 1: Fractal ViT architecture. Left: Attention diagram. For clarity, only the attention among different types of tokens at the lower-right corner is drawn as arrows. The toy example shown here uses 2-summary that assigns a summary token for every 2 × 2 regular tokens. Right: Regular tokens created from linear projection of RGB values of patches are fed to the transformer encoder along with zero-init summary tokens and global token, optionally after adding the positional encoding.

2024) or 4-summary that assigns a summary token for every 4×4 regular tokens), and whether fractal mask is applied. Unlike registers, summary tokens always use the same positional encoding as that of regular tokens (sincos2d, learned, 2D-ALiBi, or none). With 256 input resolution and patch size 16, there are 4×4 summary tokens and one global token, so the total number of tokens is constant among the variants. We can see that sincos2d outperform both learned positional encoding and 2D-ALiBi, while models without positional encoding perform the worst. Positional encoding of the additional tokens turns out to be inconsequential and fractal mask doesn't improve model performance, even for models without positional encoding.

With fractal mask shown to be unhelpful, we decide to remove it and further test whether registers/-summary tokens with different positional encodings are useful at all. Since we are no longer limited to using powers of 4 as the number of regular tokens, we revert to the standard 224 input resolution with the best-performing sincos2d positional encoding (Table 2). Here k-summary again means adding a summary token for every $k \times k$ tokens. When k is small enough ($k = 2$), we create "summary of summary tokens" as long as doing so results in nonzero additional tokens. When k^m doesn't divide the input resolution, we take the floor. The number of additional tokens is again controlled as we compare ViT with registers to ViT with summary tokens, e.g. $\lfloor \frac{14}{2} \rfloor^2 + \lfloor \frac{14}{2^2} \rfloor^2 + \lfloor \frac{14}{2^3} \rfloor^2 = 59$. As we can see, additional tokens do not improve model performance.

Table 1: Top-1 validation set accuracy of variants of positional encoding, additional tokens, and fractal mask.

Variant	Top-1 val acc.
sincos2d + 17 registers	**77.68**
sincos2d + 4-summary	77.61
sincos2d + 4-summary + fractal mask	77.57
learned + 4-summary	76.63
learned + 4-summary + fractal mask	76.11
2D-ALiBi + 4-summary	76.16
2D-ALiBi + 4-summary + fractal mask	76.26
none + 17 registers	72.93
none + 4-summary + fractal mask	73.09

Table 2: Top-1 validation set accuracy with different number of registers/summary tokens.

Variant	Top-1 validation set accuracy
59 registers	77.06
17 registers	77.08
9 registers	**77.12**
4 registers	77.07
2-summary	77.10
3-summary	77.01
4-summary	77.05
5-summary	77.02

5 Conclusion

The ImageNet-1k ViT-S/16 baseline remains unbeaten: Adding registers or summary tokens do not improve it and neither do fractal masks, in isolation or in combination with various positional encodings. Perhaps it is imperative to reexamine the studies that inspired these experiments:

1. (Darcet et al., 2024) reports that outlier tokens with large norm only appear in larger models. Specifically for ViT, they are found to appear in models ”larger than and including ViT-Large” for the common ViT sizes. Since registers are meant to mitigate outlier tokens, perhaps it is not surprising that they do not help smaller models like ViT-S/16.
2. Transformers without positional encoding (NoPE, Kazemnejad et al. (2023)) that fully relies on the causal mask for breaking permutation invariance and inspired the fractal mask were only tested in the domain of language models. Furthermore, the mask for NoPE applies to tokens directly tied to the input (token embeddings) instead of dedicated attention sink (Xiao et al., 2024) or additional tokens like registers.
3. CROMA (Fuller et al., 2024) uses 2D-ALiBi and shows that it outperforms sincos2d in ablation but we do not find it advantageous for the ImageNet-1k ViT-S/16 baseline. Overall CROMA is a distantly-related model for a different purpose (multimodal representation learning for satellite images) so the qestion on what makes the difference remains open. However, perhaps the reasons for 2D-ALiBi’s performance in CROMA offered in (Fuller et al., 2024) still hold true:
   (a) Satellite imagery is rotation-invariant so it is a better match for 2D-ALiBi, which at least exhibits symmetry of dihedral group D_4. In contrast, ImageNet is at most symmetric with respect to horizontal flip.
   (b) 2D-ALiBi limits attention weights of distant tokens and helps avoid representational collapse due to contrastive objectives, which is not applicable for the ImageNet-1k ViT-S/16 baseline.

Based on these comparisons, we believe that the following future directions may be worth pursuing:

1. It may be worth testing fractal masks and summary tokens with different positional encodings again at the scale when outlier tokens become a problem and registers start to help.
2. Especially in light of the phenomenon of attention sink in Large Language Models (LLMs), it may be worth bringing the idea of fractal masks and summary tokens back to the domain of language models and see if they improve the performance of decoder models or masked encoder models.
3. Positional encodings may need to be customized to respect the underlying symmetries of the input for better performance. In fact, for the application of satellite imagery, it may be worth trying architectures that fully respect $E(2)$ symmetry such as (Xu et al., 2023) instead of 2D-ALiBi that merely exhibits symmetry of D_4.

REFERENCES

Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. Better plain vit baselines for imagenet-1k. *arXiv preprint arXiv:2205.01580*, 2022.

Tianlong Chen, Yu Cheng, Zhe Gan, Lu Yuan, Lei Zhang, and Zhangyang Wang. Chasing sparsity in vision transformers: an end-to-end exploration. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021a. Curran Associates Inc. ISBN 9781713845393.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9620–9629, 2021b. doi: 10.1109/ICCV48922.2021.00950.

Xuanyao Chen, Zhijian Liu, Haotian Tang, Li Yi, Hang Zhao, and Song Han. Sparsevit: Revisiting activation sparsity for efficient high-resolution vision transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2061–2070, 2023. doi: 10.1109/CVPR52729.2023.00205.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=2dnO3LLiJ1`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Anthony Fuller, Koreen Millard, and James Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36, 2024.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1382–1390, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.99. URL `https://aclanthology.org/2022.findings-emnlp.99/`.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pp. 289–305. Springer, 2025.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL `https://doi.org/10.5281/zenodo.5143773`.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=Drrl2gcjzl`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=a68SUt6zFt`. Featured Certification.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

TimDarcet @TimDarcet. Thanks! 1. the only diff is that the [cls] is returned by the forward, while the registers are discarded. no difference in the way the parameters are declared & init / the way they are added to the sequence at the beginning of the forward. `https://twitter.com/TimDarcet/status/1707821795837952100`, 9 2023. Accessed: 2025-01-31.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of vit. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pp. 516–533, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20052-6. doi: 10.1007/978-3-031-20053-3_30. URL `https://doi.org/10.1007/978-3-031-20053-3_30`.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=NG7sS51zVF`.

Renjun Xu, Kaifan Yang, Ke Liu, and Fengxiang He. $e(2)$-equivariant vision transformer. In Robin J. Evans and Ilya Shpitser (eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2356–2366. PMLR, 31 Jul–04 Aug 2023. URL `https://proceedings.mlr.press/v216/xu23b.html`.

A 4-SUMMARY MASK SNIPPET

Here is a working function for creating a 4-summary mask in Python and PyTorch. The implementation we use in production is more general but more complicated.

```python
def create_fractal_attention_mask(n_h, n_w):
    # Create mask for regular tokens, summary tokens, and global token
    mask_16x16, mask_4x4, mask_global = torch.ones(n_h * n_w, n_h * n_w), torch.ones(n_h * n_w // 16, n_h * n_w // 16), torch.ones(1, 1)

    # Combine masks
    mask = torch.block_diag(mask_16x16, mask_4x4, mask_global)

    # Allow 4x4 summary tokens to attend to their corresponding 4x4 regions and vice versa
    for i in range(n_h // 4):
        for j in range(n_w // 4):
            index = n_h * n_w + i * 4 + j
            for row in range(i * 4, i * 4 + 4):
                start = row * n_w + j * 4
                mask[start:start + 4, index] = mask[index, start:start + 4] = 1

    # Allow global token to attend to everything
    mask[-1, :] = mask[:, -1] = 1
    return mask
```

Listing 1: 4-summary mask snippet

B Positional Encoding

Let us define the following notations:

- $\mathbf{e}_i \in \mathbb{R}^d$ denotes the patch embedding for position i
- $d \in \mathbb{N}$ represents the embedding dimension
- $\boldsymbol{\pi}_i \in \mathbb{R}^d$ represents the learned position embedding for position i
- $\mathbf{t}_i \in \mathbb{R}^d$ denotes the final token representation at position i

The positional encoding mechanism maintains spatial information within the transformer architecture. Traditional Vision Transformers employ learned position embeddings $\boldsymbol{\pi}_i \in \mathbb{R}^d$, which are added to the patch embeddings to form tokens:

$$\mathbf{t}_i = \mathbf{e}_i + \boldsymbol{\pi}_i \tag{1}$$

B.1 ALiBi: Position-Aware Attention Mechanism

Let us define:

- $\mathbf{q}_i \in \mathbb{R}^d$ denotes the query vector at position i
- $\mathbf{k}_j \in \mathbb{R}^d$ denotes the key vector at position j
- $h \in \{0, 1, ..., n-1\}$ represents the attention head index, assuming that we have n attention heads
- $m(h) \in \mathbb{R}^+$ is a head-specific slope parameter
- $a_{ij}^h \in \mathbb{R}$ represents the attention score between positions i and j for head h

The ALiBi mechanism modifies attention computation through position-dependent biases. For attention head h, the attention score between positions i and j is defined as:

$$a_{ij}^h = \frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d}} - m(h) \cdot |i - j| \tag{2}$$

where $m(h)$ follows the geometric sequence:

$$m(h) = 2^{-\frac{8(h+1)}{n}}, \quad h \in \{0, 1, ..., n-1\} \tag{3}$$

B.1.1 2D ALiBi Extension

For the 2D variant, we define:

- $(i, j), (k, l) \in \mathbb{N}^2$ represent spatial coordinates in the 2D grid
- $\mathbf{q}_{(i,j)} \in \mathbb{R}^d$ denotes the query vector at position (i, j)
- $\mathbf{k}_{(k,l)} \in \mathbb{R}^d$ denotes the key vector at position (k, l)

The attention score is computed as follows:

$$a_{(i,j),(k,l)}^h = \frac{\mathbf{q}_{(i,j)}^\top \mathbf{k}_{(k,l)}}{\sqrt{d}} - m(h) \cdot \sqrt{(i-k)^2 + (j-l)^2} \tag{4}$$

B.2 SINUSOIDAL 2D POSITIONAL EMBEDDINGS

Let us define:

- $h, w \in \mathbb{N}$ represent the height and width of the feature map
- $PE_{(y,x,i)} \in \mathbb{R}$ denotes the positional encoding at spatial position (y, x) for dimension i
- $\omega_i \in \mathbb{R}^+$ represents the frequency for dimension i
- temperature $\tau \in \mathbb{R}^+$ is a hyperparameter controlling the frequency spectrum

For a feature map of dimensions $h \times w$, the position encoding is computed as:

$$PE_{(y,x,2i)} = \sin(y\omega_i), \quad PE_{(y,x,2i+1)} = \cos(y\omega_i) \tag{5}$$

$$PE_{(y,x,2i+d/2)} = \sin(x\omega_i), \quad PE_{(y,x,2i+d/2+1)} = \cos(x\omega_i) \tag{6}$$

where $\omega_i = \tau^{-4i/d}$ determines the frequency for dimension i.

These formulations ensure unique spatial position encoding while maintaining relative positional relationships across scales.