Vision Transformer Architectures with Registers (Thesis outline)

Florian Weidner
Philipps-University Marburg, Germany
Department of Mathematics and Computer Science, Deep Learning Group
February 09, 2024

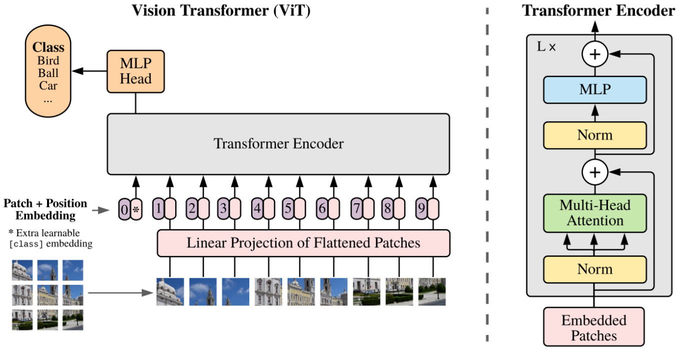

Fig. 1. Overview of a ViT architecture. [2]

***Abstract*—The abstract goes here.**

***Index Terms*—Vision Transformer (ViT),**

I. Introduction

Introduction to the topic...

Explanation of ViTs [4] [2] [**ruan2022visiontransformersstateart**] [5]

II. Vision Transformers

The Transformer architecture is a neural network model architecture, created primarily for sequence-to-sequence tasks in Natural Language Processing (NLP). It consists of an encoder, which makes the input sequence into a continuos representation and a decoder, which then generates the output sequence. The encoder is built up of n identical layers, containing following components:

- multi-head self-attention mechanism: captures relationships between all tokens in the input, regardless of their distance
- feed-forward network: simple two-layer MLP network with ReLU activation which is applied to each token separately
- add & norm layers using residual connections and layer normalization to stabilize the training

The result outcome of the encoder is a enriched sequence representation, which is then used by the decoder to generate the output sequence. The decoder also consits of n identical layers with:

- masked multi-head self-attention mechanism: ensures a causal generation, by preventing that tokens have impact to future tokens
- encoder-decoder attention mechanism: focuses on the relevant parts of the encoder's output
- feed-forward network: similar to the encoder
- add & norm layer:ssimilar to the encoder

The input text is embedded and combined with a positional encoding to provide token order information. Because several attention layers can run in parallel, the architecture is significantly more parappelizable than Recurrent Neural Network (RNN) or Convolutional Neural Network (CNN) architectures, which makes it very efficient for modern hardware accelerators. That allows the Transformer to scale to very large models and datasets. [9]

Dosovitskiy et al. introduced the idea of using the stated transformer architecture for computer vision. A lot of research tried to combine self-attention mechanisms with CNN architectures, not achieving a effectively scalable method for modern hardware accelerators. [2] proposed to apply a standard Transformer directly to images, that are split into fixed-size patches. Each patch is flattened into a vector and passed through a linear projection layer to form an embedding as input for the Transformer. These embeddings are used as tokens in a NLP scenario. Positional embeddings are added to retain spatial information since they process images as sequences, unlike CNNs which inherently capture spatial hierarchies. For classification tasks, an extra learnable [class] embeding is added in front of the embedded input. At the output of the encoder, the final representation of this token is used for classification. ViTs have much less image-specific inductive bias than CNNs, because other than CNNs, with the global self-attention mechanism spatial relationships needs to learned from scratch, but long-range dependencies across the entire image can be captured. As Transformers, ViTs are normally pre-trained on large datasets and then fine-tuned to more specific tasks. After pre-training, the prediction head is removed and a zero-initialized feedforeward layer,where the size is the number of classes, is added. Like Transformers, ViTs are also very parallelizable, which makes them very efficient. But [2] found out that without large-scale pre-training, ViTs often underperform. So ViTs requires sidnificant computational resources. But when pre-trained on large datasets, ViTs outperforms CNNs on image classification tasks. The architecture performs well for transfer learning, where the pre-trained model can be fine-tuned already with limited labeled data. [2] [2] stated that further scaling of ViT would likely lead to improved performance. Alo self-supervised pre-training can be improved. They found out that with mimicking the masked language modeling task used in BERT, the model performs still better than CNNs but a bit worse that with supervised pre-training. By now different architectures and training-tricks of ViTs have been proposed to further improve ViTs including self-supervised learning.

The architecture got adapted for image recognition, object detection, image segmentation, pose estimation, and 3D reconstruction tasks. [6]

III. Vision Transformers need registers: A summary

Summerization and explanation of the paper [1] [10]

IV. Comparison to other papers with performance improvements of ViTs

survey: [3]

I thought of comparing the previous explained paper [1] with papers like [11] [12] [8] [7] [5] which try to improve efficiency of ViTs

V. Conclusion

The conclusion goes here.

References

[1] Timothée Darcet et al. *Vision Transformers Need Registers*. 2024. arXiv: 2309.16588 [cs.CV]. URL: https://arxiv.org/abs/2309.16588.

[2] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV]. URL: https://arxiv.org/abs/2010.11929.

[3] Quentin Fournier, Gaétan Marceau Caron, and Daniel Aloise. "A Practical Survey on Faster and Lighter Transformers". In: *ACM Comput. Surv.* 55.14s (July 2023). ISSN: 0360-0300. DOI: 10.1145/3586074. URL: https://doi.org/10.1145/3586074.

[4] Salman Khan et al. "Transformers in Vision: A Survey". In: *ACM Comput. Surv.* 54.10s (Sept. 2022). ISSN: 0360-0300. DOI: 10.1145/3505244. URL: https://doi.org/10.1145/3505244.

[5] Yun Liu et al. "Vision transformers with hierarchical attention". en. In: *Mach. Intell. Res.* 21.4 (2024), pp. 670–683.

[6] Bo-Kai Ruan, Hong-Han Shuai, and Wen-Huang Cheng. *Vision Transformers: State of the Art and Research Challenges*. 2022. arXiv: 2207.03041 [cs.CV]. URL: https://arxiv.org/abs/2207.03041.

[7] Michael S. Ryoo et al. *TokenLearner: What Can 8 Learned Tokens Do for Images and Videos?* 2022. arXiv: 2106.11297 [cs.CV]. URL: https://arxiv.org/abs/2106.11297.

[8] Hugo Touvron, Matthieu Cord, and Hervé Jégou. "DeiT III: Revenge of the ViT". In: *Computer Vision – ECCV 2022*. Ed. by Shai Avidan et al. Cham: Springer Nature Switzerland, 2022, pp. 516–533. ISBN: 978-3-031-20053-3.

[9] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

[10] Feng Wang et al. *Mamba-R: Vision Mamba ALSO Needs Registers*. 2024. arXiv: 2405.14858 [cs.CV]. URL: https://arxiv.org/abs/2405.14858.

[11] Yuxin Wen et al. *Efficient Vision-Language Models by Summarizing Visual Tokens into Compact Registers*. 2024. arXiv: 2410.14072 [cs.CV]. URL: https://arxiv.org/abs/2410.14072.

[12] Hongxu Yin et al. "A-ViT: Adaptive Tokens for Efficient Vision Transformer". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 10809–10818.