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

I. Preliminary structure

1) Abstract
2) Introduction
3) Vision Transformer (ViT) need Registers
4) Comparison to papers with performance improvements of ViTs
5) Conclusion

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

II. Introduction

Introduction to the topic...
Explanation of ViTs [4] [2] [6] [5]

III. Using Transistors in ViTs networks

Summerization and explanation of the paper [1] [9]

IV. Comparison to papers with performance improvements of ViTs

survey: [3]
I thought of comparing the previous explained paper [1] with papers like [10] [11] [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] Feng Wang et al. *Mamba-R: Vision Mamba ALSO Needs Registers*. 2024. arXiv: 2405.14858 [cs.CV]. URL: https://arxiv.org/abs/2405.14858.

[10] 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.

[11] 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.