Early diagnosis of skin cancer from phone-taken skin lesion images using Vision Transformers

Sina Garazhian[*1] SINA.GARAZHIAN@MPINAT.MPG.DE
[1] *Research group of Quantitative and Systems Biology, Max-Planck-Institute for Multidisciplinary Sciences (MPI-NAT), University of Göttingen, Göttingen, Germany*

Parsa Hariri[*2]
HARIRI.PARSA@HELMHOTLZ-MUNICH.DE
[2] *Computational Discovery Research Group, Institute of Computational Biology, Helmholtz Zentrum München – German Research Center for Environmental Health (GmbH), Neuherberg, Germany*
Department of Life Sciences, Technical University of Munich (TUM), Munich, Germany

Editors: Under Review for MIDL 2025

Abstract

Recent advances in computer vision have made Vision Transformers (ViTs) strong alternatives to CNNs in medical imaging. We compare top ViT models—including Token-to-Token ViT, CaiT, LeViT, ATSViT, and XCiT—on the Kaggle skin cancer dataset, focusing on classification accuracy, real score, and model complexity. While ViTs for small datasets show high accuracy, they have many parameters; LeViT offers strong performance with the fewest parameters. This review highlights current trends, deployment challenges, and future directions for transformers in skin cancer detection.

Keywords: Vision Transformers, Deep Learning, Skin Cancer, Dermatology images.

1. Introduction

Transformers, first designed for NLP tasks, gained popularity through models like BERT and RoBERTa (Vaswani et al., 2017; Devlin et al., 2019; Liu et al., 2019). Their success led to applications in computer vision (CV), where CNNs had traditionally dominated (He et al., 2016a,b; Tan and Le, 2019). Early ViTs combined attention with convolution (Bello et al., 2019), but newer versions rely solely on self-attention.

ViTs have since been applied to image classification (Dosovitskiy et al., 2020; Touvron et al., 2021a), segmentation (Ye et al., 2019a), object detection (Ye et al., 2019b), and video analysis (Sun et al., 2019). The original Vision Transformer (Dosovitskiy et al., 2020) showed pure transformer models could excel in CV, inspiring further research. Studies (Azad et al., 2024; Liu et al., 2023) explored ViTs in medical imaging, and (Khalil et al., 2023) reviewed their evolution into lighter, efficient models.

While most reviews focus on clinic-acquired images from mid-to-late cancer stages, our review evaluates recent ViTs for early skin cancer detection using phone-quality lesion images.

[*] Contributed equally

2. Dataset

We used the ISIC 2024 Kaggle dataset (Kurtansky et al., 2024), which includes 401K 3D Total Body Photography (TBP) images mimicking non-dermoscopic photos. Captured with the Vectra WB360, the images cover the full skin surface. AI software detects and crops individual lesions into 15×15 mm images.

3. Results

Our results in Figure 1 highlight a non-linear relationship between model complexity and real-world performance. The Vision Transformer for Small Datasets (Lee et al., 2021) achieved the highest real score of 132, indicating superior performance and generalization by incorporating Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA) which increases the receptive field during tokenization and sharpens attention scores, respectively. Notably, it did so with a moderate parameter count (∼54M) and 50 training epochs, showcasing that well-designed, domain-adapted ViTs can outperform larger architectures when carefully tuned for small-scale medical datasets. Despite having the same accuracy (92%) as several other models, its higher real score suggests better optimization and convergence behavior over training.

In contrast, CaiT (Touvron et al., 2021b) which utilizes LayerScale, a learnable per-channel residual scaling mechanism that facilitates the training of deep transformers, the largest model in our study with over 120 million parameters, underperformed significantly with a real score of 97, the lowest among all models including the plain ViT model (Beyer et al., 2022). Although it reached a marginally higher accuracy (93%), its short training duration (10 epochs) likely hindered its potential. This illustrates the importance of not only model capacity but also sufficient training time for transformers to fully utilize their representational power.

LeViT (Graham et al., 2021) which employs a multi-stage transformer design incorporating CNN-like components, stands out as the most efficient model in our benchmark, achieving a real score of 125 and the highest classification accuracy (94%) with a remarkably small footprint of just 17M parameters. This model is particularly well-suited for real-time or embedded diagnostic applications, where computational resources are limited. The result also confirms the efficacy of hybrid convolution-attention designs in achieving competitive performance with minimal complexity. Token-to-Token ViT (Yuan et al., 2021) which improves the tokenization process by recursively aggregating neighboring tokens, preserving local structure through a Tokens-to-Token transformation, and ats ViT (Fayyaz et al., 2022) which introduces a parameter-free Adaptive Token Sampler (ATS) module that dynamically selects informative tokens per input image. This adaptivity allows for significant reduction in token count and Giga Floating Point Operations per second (GFLOPs) during inference, each scoring 113 in real score with comparable accuracies (92%), illustrate the potential of patch re-encoding and attention-based scaling for performance gains. However, the Token-to-Token ViT model encountered memory issues during training, emphasizing a practical limitation despite its otherwise balanced architecture and low parameter count (∼20M).

Emerging architectural innovations such as patch merging and cross-covariance attention also showed strong results. The xcit (El-Nouby et al., 2021) model which proposes a novel Cross-Covariance Attention (XCA) mechanism that operates across feature channels instead

of tokens with real score of 122, ∼12M parameters and vit_with_patch_merger (Renggli et al., 2022) (real score: 118, params: ∼77M) which incorporates a lightweight module that merges redundant tokens between transformer layers, both performed well under limited training epochs (25), indicating strong inductive biases and fast convergence capabilities. Their performance suggests that such design choices may significantly enhance model efficiency and should be considered in future ViT developments for medical imaging tasks.

All models, except for Token-to-Token ViT, were successfully trained under the same computational environment. Most models converged well within 25 to 50 epochs, with higher epoch budgets yielding more stable learning curves (e.g., LeViT with 100 epochs). Models trained with fewer epochs (like CaiT and xcit) exhibited more variability in performance, reinforcing the need for longer training schedules especially for larger architectures.

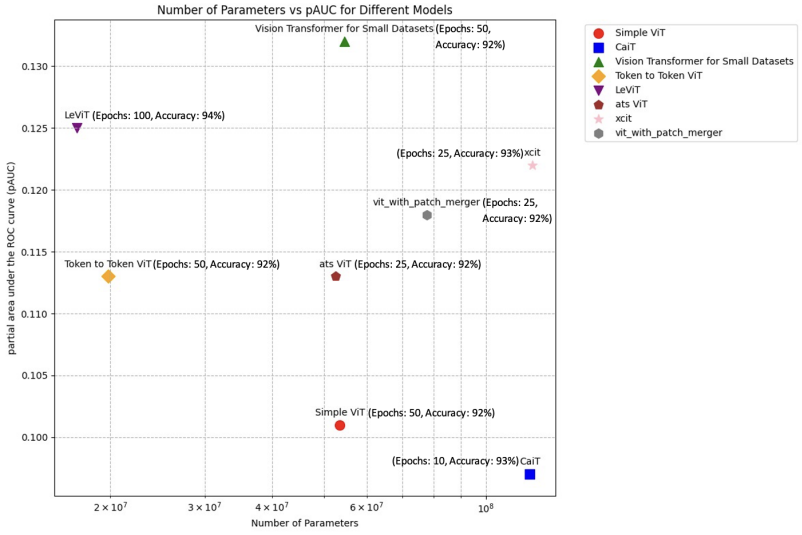

Figure 1: Models performance namely real score were described as partial area under the ROC curve (pAUC) above 80% true positive rate (TPR) since the TPR below 80% is unacceptable in clinical practice.

References

Reza Azad, Amirhossein Kazerouni, Moein Heidari, Ehsan Khodapanah Aghdam, Amirali Molaei, Yiwei Jia, Abin Jose, Rijo Roy, and Dorit Merhof. Advances in medical image analysis with vision transformers: a comprehensive review. *Medical Image Analysis*, 91: 103000, 2024.

Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3286–3295, 2019.

Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. Better plain vit baselines for imagenet-1k, 2022. URL https://arxiv.org/abs/2205.01580.

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)*, pages 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jegou. Xcit: Cross-covariance image transformers, 2021. URL https://arxiv.org/abs/2106.09681.

Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Juergen Gall. Adaptive token sampling for efficient vision transformers, 2022. URL https://arxiv.org/abs/2111.15667.

Ben Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference, 2021. URL https://arxiv.org/abs/2104.01136.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016b.

Mahmoud Khalil, Ahmad Khalil, and Alioune Ngom. A comprehensive study of vision transformers in image classification tasks. *arXiv preprint arXiv:2312.01232*, 2023.

Nicholas Kurtansky, Veronica Rotemberg, Maura Gillis, Kivanc Kose, Walter Reade, and Ashley Chow. Isic 2024 - skin cancer detection with 3d-tbp. https://kaggle.com/competitions/isic-2024-challenge, 2024. Kaggle.

Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets, 2021. URL https://arxiv.org/abs/2112.13492.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Zhaoshan Liu, Qiujie Lv, Ziduo Yang, Yifan Li, Chau Hung Lee, and Lei Shen. Recent progress in transformer-based medical image analysis. *Computers in Biology and Medicine*, 164:107268, 2023.

Cedric Renggli, André Susano Pinto, Neil Houlsby, Basil Mustafa, Joan Puigcerver, and Carlos Riquelme. Learning to merge tokens in vision transformers, 2022. URL https://arxiv.org/abs/2202.12015.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021a.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers, 2021b. URL https://arxiv.org/abs/2103.17239.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511, 2019a.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511, 2019b.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet, 2021. URL https://arxiv.org/abs/2101.11986.