WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 4, Number 2, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v4n2.37
Journal homepage: https://wepub.org/index.php/IJCSIT/index

A Brief Review of Lightweighting Methods for Vision Transformers (ViT)

Xiang Du

Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, 210023, China
P21000613@njupt.edu

ABSTRACT

The Vision Transformer (ViT) has emerged as a powerful model in recent years, surpassing traditional Convolutional Neural Networks (CNNs) in various benchmarks. However, its large model architecture and high parameter count present significant challenges for mobile deployment. Consequently, the exceptional performance of ViT in computer vision tasks is often overshadowed by its difficulties in being deployed on mobile devices due to its large parameter size and high computational demands. This paper provides a comprehensive review of the literature on lightweight ViT models, focusing on model optimization strategies such as post-training modifications—quantization, pruning, and knowledge distillation—as well as architectural changes including hybrid CNN-Transformer, MLP-based, and sparse models. These strategies are aimed at improving efficiency for mobile platforms. The review aims to clarify current techniques for mobile ViT, guide future research, stimulate innovation, and contribute to the development of efficient ViT models for mobile environments.

KEYWORDS

Vision Transformer; Lightweight Strategies; Mobile Deployment; Post-Training Modifications; Model Architecture Changes

1. INTRODUCTION

In the realm of deep learning, Convolutional Neural Networks (CNNs) have become a staple on mobile devices, renowned for their high precision and swift response in tasks such as image recognition and object detection. MobileNetV1and its successor, MobileNetV2, have notably minimized model size and computational demands through the innovative use of Depthwise Separable Convolution (DSC), all while sustaining high levels of accuracy [1-3]. However, the introduction of the Vision Transformer (ViT) in 2020 has disrupted the status quo. ViT, bolstered by its novel Multi-Headed Self-Attention (MHSA) and Multi-Layer Perceptron (MLP) components, has demonstrated the capacity to outperform CNNs in several visual tasks [4]. A case in point is its performance on the ImageNet dataset, where it has secured a Top-1 accuracy between 77.9% and 81.3% [5].

Despite the Vision Transformer's (ViT) impressive accuracy, its extensive model size and computational heft present a barrier to its implementation on mobile devices with limited resources [6]. The MHSA operation in ViT incurs a quadratic complexity, and the MLP module's process of expanding the embedding space, applying nonlinear transformations, and then projecting back to the original dimensions, exacerbates inference latency [7]. Consequently, while ViT outshines CNNs in terms of performance, it experiences 1.5 to 4 times the inference latency compared to CNNs [8, 9]. Despite its broad applications in visual recognition tasks such as classification, detection, and

segmentation, CNNs remain the go-to architecture for mobile devices due to their inference efficiency [5, 10-14]. To bridge this gap, researchers are delving into lightweight adaptations of ViT, seeking to curtail model size and accelerate operational speed without compromising on accuracy. This paper delineates the lightweighting strategies, encompassing post-training model modifications and architectural innovations [15].

2. POST-TRAINING MODIFICATIONS

Post-training model modifications encompass a suite of techniques aimed at enhancing the efficiency of Vision Transformers (ViT) on mobile devices. These include quantization, which compresses the model by reducing the precision of its parameters, thereby accelerating inference [16, 17]; pruning, which refines the model by eliminating less significant weights [2]; and knowledge distillation, a process that imparts the learnings of a sophisticated model to a more compact one, enhancing its performance [18]. Collectively, these strategies are leveraged to optimize ViT for deployment on resource-constrained platforms.

2.1. Quantization

The implementation of quantization in Vision Transformers (ViTs) targets pivotal parameters like weights, activations, and attention matrices. This process, which translates floating-point numbers into integers with reduced bit-width, substantially diminishes the model's storage needs and computational load. In the case of I-VIT, an integer-only quantization approach enables ViTs to conduct the full inference cycle without any reliance on floating-point operations, leading to a significant reduction in model size and a substantial 3.72 to 4.11-fold speedup in inference on GPUs with TVM deployment [19]. Furthermore, the PTQ4ViT model, which identifies the optimal scaling factor through a bi-uniform quantization and Hessian-guided metric tailored to the activation value distribution in ViTs, experiences a minimal Top-1 accuracy drop of less than 0.5% when subjected to 8-bit quantization on the ImageNet dataset [20]. This illustrates how quantization can be instrumental in sustaining high accuracy while optimizing memory usage and computational resources. These studies not only facilitate the adoption of ViTs on devices with limited resources but also offer novel insights and methodologies for the efficient inference of deep learning models in practical scenarios.

2.2. Pruning

Optimization of Vision Transformers (ViTs) through pruning techniques has led to notable enhancements in performance and efficiency. Sparse attentional pruning pinpoints and eliminates weights with minimal impact on the model's output by scrutinizing the trained attentional weights. For instance, the LeIAP algorithm in the FormerLeaf model streamlines the Transformer's attentional heads to expedite inference and shrink the model's footprint [21]. Structured pruning, as seen in the SPViT model, efficiently trims the ViT by sharing parameters between Multi-Head Self-Attention (MSA) and convolutional operations, employing a single-path search framework that automatically determines the conversion of MSA layers to convolutional layers and the optimal dimensions for Feed-Forward Network (FFN) layers [22]. Iterative pruning, exemplified by the NViT model, furthers the cause through global structural pruning guided by the Hessian matrix and delay-aware regularization, enhancing the model's operational velocity and reducing reference counts while preserving accuracy [23]. Empirical assessments on the Cassava Leaf Disease Dataset reveal that the pruned FormerLeaf model significantly reduces model size, accelerates evaluation, and enhances accuracy by approximately 3%, with a 10% reduction in training time. These innovations not only bolster the ViT's computational efficiency but also offer streamlined solutions for model storage and deployment in resource-limited settings, enabling the pruned model to conduct inference with compact size, swift speed, and sustained high accuracy.

2.3. Knowledge Distillation

In the domain of Vision Transformer (ViT) knowledge distillation, researchers have delved into diverse strategies to bolster the models' efficiency and efficacy. Knowledge distillation is primarily bifurcated into two categories: soft target distillation and feature distillation. Soft target distillation concentrates on replicating the output distribution of extensive models by minimizing the cross-entropy between the models' outputs. This method in ViTs aids compact models in grasping the decision-making demarcations of their larger counterparts, thereby enhancing classification precision. An exemplar of this is the attention mechanism-based distillation proposed by Touvron et al. [24]. Conversely, feature distillation zeroes in on capturing feature representations from the intermediate layers of a large model. A case in point is the Unified Visual Transformer compression framework introduced by Yu et al., which amalgamates structured pruning, jump-join configuration, and knowledge distillation [25]. Furthermore, Wang et al. put forth an attentional probe-based distillation method [26]. This technique leverages valuable data sifted from an abundance of unlabeled data to train a nimble student Transformer. It maximizes the alignment between a robust teacher model and a compact student model concerning both output and intermediate features via a probe knowledge distillation algorithm. By harnessing the synergies of soft targets and feature distillation, this methodology empowers ViTs to rival the performance of large models, all the while retaining a modest model size.

3. MODEL ARCHITECTURE CHANGES

Model architecture modifications for Vision Transformers (ViTs) are tailored to suit the constraints of mobile devices. This adaptation includes the development of hybrid models that integrate Convolutional Neural Networks (CNNs) with Transformers. An example is MobileViT, which enhances local feature extraction by incorporating a CNN module within the ViT framework [8]. Additionally, MLP-based models like EfficientForme streamline model complexity by substituting certain ViT components with Multi-Layer Perceptrons (MLPs) [27]. Furthermore, the advent of sparse models addresses computational demands by inducing sparsity within the model, achieved through techniques such as structured pruning or weight sparsification [28, 29]. These architectural innovations aim to optimize ViTs for mobile deployment, balancing performance with the efficiency required for resource-limited environments.

3.1. CNN Hybrid Models

The pursuit of lightweight Vision Transformer (ViT) models that integrate Convolutional Neural Networks (CNNs) is driven by two key considerations: structural design and the attention mechanism, both aimed at enhancing the model's efficacy on devices with limited resources. Structurally, these models blend the CNN's adeptness at local feature extraction with the Transformer's prowess in handling global information. A case in point is LeViT , which, by integrating attentional bias and a multi-stage Transformer architecture, manages to offer swift inference speeds and high accuracy through reduced feature dimensions and efficient Patch descriptor computation [30]

Building upon this, RepViT merges the architectural design of lightweight ViTs with MobileNetV3, enhancing mobile performance through incremental improvements, achieving over 80% top-1 accuracy on ImageNet and an impressive 1.0 ms latency on an iPhone 12 [31].

Furthermore, enhancing the attention mechanism is pivotal for boosting the lightweight ViT's performance. By refining the self-attention module or introducing novel attention forms, these models can more efficiently capture global contextual cues. ResViT exemplifies this with its encoder-decoder architecture, embedding Aggregate Residual Transformer (ART) blocks that amalgamate CNN and Transformer modules, thereby enhancing the model's capacity to learn long-range spatial dependencies, particularly excelling in tasks involving multimodal medical image synthesis [32].

3.2. MLP-based Models

The Vision Transformer (ViT) lightweight models that leverage Multi-Layer Perceptrons (MLPs) have made significant strides in enhancing efficiency and reducing parameters. They achieve this by strategically deploying MLP modules at pivotal junctures, such as:

(1) Replacing the Self-Attention Layer: This modification diminishes the computational expenditure associated with traditional ViT architectures.

(2) Simplifying the Feed-Forward Network (FFN): By streamlining the FFN, the number of parameters is curtailed.

(3) Enhancing the Embedding Layer: Improvements here allow for a more efficient processing of image patches.

The ResMLP model exemplifies this approach with a fully MLP-based architecture [33]. It projects image patches through linear layers and employs two types of residual operations for representation updates: a cross-patch linear layer and an MLP across channels. This streamlined design eschews the self-attention layer and the intricate FFN structure, demonstrating commendable accuracy on the ImageNet dataset. For instance, the ResMLP-S12 model garnered a Top-1 accuracy of 76.6%, and the ResMLP-B24 model achieved 81.0% accuracy without the need for additional data.

The CycleMLP model introduces the Cycle Fully-Connected Layer (Cycle FC), an innovative fully connected layer. Each Cycle FC block encompasses a channel-MLP, which is composed of two linear layers interspersed with a GELU (Gaussian Error Linear Unit) activation function [34]. The channel-MLP operates between Cycle FC layers to refine feature processing in the channel dimension. Cycle FC expands the model's receptive field by applying distinct sampling offsets on the feature map, sampled periodically at a set stride, thus avoiding an increase in computational load. This design enables the model to manage inputs of varying resolutions and bolsters its capacity for spatial context aggregation. On the ADE20K dataset for instance segmentation tasks, the CycleMLP-Tiny model secured an mIoU of 81.6% with a lower computational footprint (1.0 ms latency), surpassing the Swin-Tiny model's mIoU of 80.3%.

3.3. Sparse Models

Sparse models are integral to optimizing neural networks by trimming down the parameter count and computational intricacy. These models sidestep conventional pruning or quantization methods, instead embracing inventive structural designs that amplify model efficiency. Take, for example, the Efficient-ViTmodel, which amalgamates Depth-Wise Convolution (DWConv) layers from MobileNetV2 with a linear attention module, efficiently expanding the receptive field without incurring extra parameter costs [35]. The DWConv layer epitomizes a structured form of sparsity, enabling independent convolution for each input channel, thus reducing the parameter count and computational load.

Furthering this approach, EdgeNeXt introduces the Split Depth-wise Transpose Attention (SDTA) encoder as an alternative to the traditional Multi-Head Self-Attention (MHSA) module [36]. The SDTA encoder partitions the input tensor into several channel groups, applying DWConv and attention mechanisms across these groups to broaden the model's receptive field while upholding parameter sparsity.

LeViT, in contrast, embraces a unique strategy by transplanting the concept of reducing the resolution of activation maps from traditional CNNs into the ViT architecture [30]. This technique minimizes the model's parameter volume and computational demands without sacrificing performance.

MobileViT , conversely, incorporates a specialized module designed to capture the interplay of local and global image features [8]. This module utilizes NxN convolutions followed by 1x1 point-wise

convolutions to seize local features, which are then channeled into a global feature capture module. Such a design empowers the model to adeptly capture and synthesize features across various scales while maintaining structured sparsity.

4. CONCLUSION

This review synthesizes lightweight strategies for Vision Transformers (ViT) on mobile devices, exploring post-training modifications and architectural innovations that enhance efficiency without sacrificing accuracy. Despite advancements, challenges in complexity reduction and performance balance persist. Future research may refine quantization, pruning, and distillation techniques, with a particular focus on integrating CNN and ViT strengths. Offers valuable insights for both academia and industry, guiding the adoption of effective lightweight approaches and advancing AI technology deployment.

REFERENCES

[1] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. ArXiv, abs/1704.04861.

[2] Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., & Chen, L. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4510-4520.

[3] Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., & Chen, L. (2018). Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation. ArXiv, abs/1801.04381.

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv, abs/2010.11929.

[5] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 9992-10002.

[6] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., & Tao, D. (2020). A Survey on Vision Transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence, PP, 1-1.

[7] Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. Neural Information Processing Systems.

[8] Mehta, S., & Rastegari, M. (2021). MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. ArXiv, abs/2110.02178.

[9] Wang, X., Zhang, L., Wang, Y., & Yang, M. (2022). Towards efficient vision transformer inference: a first study of transformers on mobile devices. Proceedings of the 23rd Annual International Workshop on Mobile Computing Systems and Applications.

[10] Caron, M., Touvron, H., Misra, I., J'egou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 9630-9640.

[11] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020) End-to-End Object Detection with Transformers, European Conference on Computer Vision, abs/2005.12872: 213-229.

[12] Li, Y., Mao, H., Girshick, R.B., & He, K. (2022). Exploring Plain Vision Transformer Backbones for Object Detection. ArXiv, abs/2203.16527.

[13] Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., & Girdhar, R. (2021). Masked-attention Mask Transformer for Universal Image Segmentation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1280-1289.

[14] Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Izacard, G., Joulin, A., Synnaeve, G., Verbeek, J., & J'egou, H. (2021). ResMLP: Feedforward Networks for Image Classification With Data-Efficient Training. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45, 5314-5321.

[15] Alam, N., Kolawole, S., Sethi, S.S., Bansali, N., & Nguyen, K. (2023). Vision Transformers for Mobile Applications: A Short Survey. ArXiv, abs/2305.19365.

[16] Han, S., Pool, J., Tran, J., & Dally, W.J. (2015). Learning both Weights and Connections for Efficient Neural Network. Neural Information Processing Systems.

[17] Han, S., Mao, H., & Dally, W.J. (2015). Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding. arXiv: Computer Vision and Pattern Recognition.

[18] Hinton, G.E., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. ArXiv, abs/1503.02531.

[19] Li, Z., & Gu, Q. (2022). I-ViT: Integer-only Quantization for Efficient Vision Transformer Inference. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 17019-17029.

[20] Yuan, Z., Xue, C., Chen, Y., Wu, Q., & Sun, G. (2021). PTQ4ViT: Post-training Quantization for Vision Transformers with Twin Uniform Quantization. European Conference on Computer Vision.

[21] Thai, H., Le, K., & Nguyen, N.N. (2023). FormerLeaf: An efficient vision transformer for Cassava Leaf Disease detection. Comput. Electron. Agric., 204, 107518.

[22] Kong, Z., Dong, P., Ma, X., Meng, X., Niu, W., Sun, M., Shen, X., Yuan, G., Ren, B., Tang, H., Qin, M., & Wang, Y. (2022) SPViT: Enabling Faster Vision Transformers via Latency-Aware Soft Token Pruning., Lecture Notes in Computer ScienceComputer Vision – ECCV 2022, 13671: 620-640.

[23] Yang, H., Yin, H., Shen, M., Molchanov, P., Li, H.H., & Kautz, J. (2021). Global Vision Transformer Pruning with Hessian-Aware Saliency. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 18547-18557.

[24] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & J'egou, H. (2020). Training data-efficient image transformers & distillation through attention. International Conference on Machine Learning.

[25] Yu, S., Chen, T., Shen, J., Yuan, H., Tan, J., Yang, S., Liu, J., & Wang, Z. (2022). Unified Visual Transformer Compression. ArXiv, abs/2203.08243.

[26] Wang, J., Cao, M., Shi, S., Wu, B., & Yang, Y. (2022). Attention Probe: Vision Transformer Distillation in the Wild. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2220-2224.

[27] sLi, Y., Yuan, G., Wen, Y., Hu, E., Evangelidis, G., Tulyakov, S., Wang, Y., & Ren, J. (2022). EfficientFormer: Vision Transformers at MobileNet Speed. ArXiv, abs/2206.01191.

[28] Gale, T., Elsen, E., & Hooker, S. (2019). The State of Sparsity in Deep Neural Networks. ArXiv, abs/1902.09574.

[29] Liu, S., & Wang, Z. (2023). Ten Lessons We Have Learned in the New "Sparseland": A Short Handbook for Sparse Neural Network Researchers. ArXiv, abs/2302.02596.

[30] Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., J'egou, H., & Douze, M. (2021). LeViT: a Vision Transformer in ConvNet’s Clothing for Faster Inference. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 12239-12249.

[31] Wang, A., Chen, H., Lin, Z., Pu, H., & Ding, G. (2023). RepViT: Revisiting Mobile CNN From ViT Perspective. ArXiv, abs/2307.09283.

[32] Dalmaz, O., Yurt, M., & Çukur, T. (2021). ResViT: Residual Vision Transformers for Multimodal Medical Image Synthesis. IEEE Transactions on Medical Imaging, 41, 2598-2614.

[33] Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Izacard, G., Joulin, A., Synnaeve, G., Verbeek, J., & J'egou, H. (2021). ResMLP: Feedforward Networks for Image Classification With Data-Efficient Training. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45, 5314-5321.

[34] Chen, S., Xie, E., Ge, C., Liang, D., & Luo, P. (2021). CycleMLP: A MLP-like Architecture for Dense Prediction. ArXiv, abs/2107.10224.

[35] Xie, Y., & Liao, Y. (2023). Efficient-ViT: A Light-Weight Classification Model Based on CNN and ViT. Proceedings of the 2023 6th International Conference on Image and Graphics Processing.

[36] Maaz, M., Shaker, A.M., Cholakkal, H., Khan, S.H., Zamir, S.W., Anwer, R.M., & Khan, F.S. (2022). EdgeNeXt: Efficiently Amalgamated CNN-Transformer Architecture for Mobile Vision Applications. ECCV Workshops.