Paper Review: FMViT — A Multiple-Frequency Mixing Vision Transformer

1. Summary of the Paper

Problem Being Addressed

The paper examines the inefficiencies of Vision Transformers (ViTs) in practical applications due to their substantial computational and memory requirements. While ViTs possess remarkable representational capabilities, they frequently exhibit inferior performance compared to conventional Convolutional Neural Networks (CNNs) in latency-critical environments, such as mobile and industrial settings. Notably, existing hybrid models that combine CNNs and Transformers have yet to establish a satisfactory balance between computational efficiency and desired performance metrics.

Main Contribution

The authors propose FMViT (Frequency-Mixing Vision Transformer), a CNN-Transformer hybrid architecture that integrates multiple innovations:

- Multi-Frequency Fusion Block (FMB) to combine high-frequency (local details) and low frequency (global context) features.
- Convolutional Fusion Block (CFB) using depthwise separable convolutions for efficient local representation learning.
- Lightweight Multi-Head Self-Attention (RLMHSA) to reduce the computational burden of self-attention while maintaining global modeling capabilities.
- gMLP, a multi-group reparameterized MLP block that improves both inference speed and accuracy.

Experimental/Theoretical Results

Experiments on standard benchmarks demonstrate FMViT's superior performance:

- Image Classification (ImageNet-1K): FMViT-L achieves 83.3% Top-1 accuracy, outperforming ResNet101 by 2.5% with equivalent latency, and matches EfficientNet-B5 performance while improving inference speed by 43%.
- Object Detection (COCO): FMViT-B achieves 3.7 APb more than ResNet101 and 16% faster inference on TensorRT.
- Semantic Segmentation (ADE20K): FMViT-B exceeds ResNet101 by 4.7 mIoU, with comparable latency on mobile and server platforms.

2. Related Work

The FMViT architecture is founded on decades of advancements in convolutional neural networks (CNNs), Vision Transformers (ViTs), and hybrid models, in conjunction with recent innovations in model reparameterization.

CNNs have been the cornerstone of computer vision since AlexNet, with architectures such as ResNet (He et al., 2016) introducing residual connections that enabled deep models without gradient vanishing. Mobile-oriented CNNs like MobileNetV1-V3 (Howard et al., 2017; Sandler et al., 2018) utilized depthwise separable convolutions to significantly reduce computational cost, making them ideal for deployment on edge devices. ShuffleNet (Zhang et al., 2018) and ConvNeXt (Liu et al., 2022) further refined this efficiency while preserving accuracy.

Transformers, initially successful in natural language processing (Vaswani et al., 2017), were introduced to computer vision by ViT (Dosovitskiy et al., 2021), which treated image patches as tokens. Despite ViT's impressive outcomes, its lack of locality and high computational burden prompted innovations such as Swin Transformer (Liu et al., 2021), which introduced hierarchical structures and shifted windows to reduce cost and preserve spatial locality. DeiT (Touvron et al., 2021) further enhanced efficiency through data-efficient training and knowledge distillation.

Hybrid models strive to combine the local precision of convolutional neural networks (CNNs) with the global context modeling capabilities of transformers. Notable examples include BoTNet (Srinivas et al., 2021), which integrates self-attention modules into CNN backbones; MobileViT (Mehta & Rastegari, 2022), which strategically blends lightweight CNNs with attention mechanisms for mobile platforms; and MobileFormer (Chen et al., 2022). EfficientFormer (Li et al., 2022d) employs neural architecture search (NAS) to discover compact yet powerful hybrid architectures.

Structural reparameterization is another relevant approach. Models are over-parameterized during training to enhance representational power, while simplified during inference to minimize latency. RepVGG (Ding et al., 2021c), ACNet (Ding et al., 2019), and DBB (Ding et al., 2021a) exemplify this technique, consolidating multiple branches into single convolutions after training. RepMLP (Ding et al., 2021b) and MobileOne (Vasu et al., 2022) further extend this concept, achieving excellent performance and low latency in mobile scenarios.

FMViT draws heavily from this body of work, particularly from NextViT (Li et al., 2022a), which similarly integrates low- and high-frequency signals. FMViT distinguishes itself through its systematic application of frequency-aware components, lightweight attention modules, and aggressive reparameterization, enabling it to achieve a superior latency-accuracy trade-off in practical deployment contexts.

In summary, FMViT builds upon an extensive body of knowledge in CNNs, ViTs, hybrid architectures, and reparameterization techniques.

Convolutional Neural Networks

- ResNet (He et al., 2016)
- MobileNetV1/V2/V3 (Howard et al., 2017; Sandler et al., 2018)
- ShuffleNet and ShuffleNetV2 (Zhang et al., 2018; Ma et al., 2018)
- ConvNeXt (Liu et al., 2022)

Vision Transformers

- ViT (Dosovitskiy et al., 2021)
- DeiT (Touvron et al., 2021)
- Swin Transformer (Liu et al., 2021)
- T2T-ViT, PiT, Reformer

Hybrid Architectures

- BoTNet (Srinivas et al., 2021)
- MobileViT and MobileFormer (Mehta & Rastegari, 2022; Chen et al., 2022)
- EfficientFormer, NextViT (Li et al., 2022)

Structural Reparameterization

- RepVGG, ACNet, DBB, RepMLP (Ding et al., 2019; 2021a; 2021b; 2021c)
- MobileOne (Vasu et al., 2022)

3. Limitations of the Paper

Manual Architecture Design: While the FMViT architecture exhibits impressive performance across image classification, object detection, and semantic segmentation, the paper presents several limitations that warrant attention. Firstly, the architecture design heavily relies on manual configuration and fixed stacking of modules, restricting the exploration of potentially more optimal configurations that could be discovered through Neural Architecture Search (NAS) or automated tuning techniques. Such methods could yield more adaptable and efficient architectures tailored to specific deployment environments.

Reproducibility: Although the authors assert industry relevance by employing TensorRT and CoreML, the reproducibility of these latency figures may be compromised due to hardware and software dependencies. Readers and practitioners without access to identical deployment environments may encounter challenges in replicating the results precisely, potentially affecting the generalizability of the claimed benefits.

Scalability: The scalability of the FMViT architecture is not comprehensively analyzed. The paper primarily evaluates mid-sized models on standard datasets, neglecting to assess FMViT’s scalability to larger datasets, higher-resolution inputs, or tasks beyond single-image processing, such as video understanding or multimodal learning.

Generalization: Fourthly, while FMViT is evaluated on well-known datasets like ImageNet-1K and COCO, its robustness and adaptability to more challenging or less curated datasets remain untested. A broader evaluation across diverse tasks and domains would corroborate its versatility.

Batch Size Considerations: The evaluation is conducted with a batch size of one, which is typical for real-time inference but not applicable to many training or high-throughput settings. Performance under varying batch sizes, particularly for server-side inference or training scenarios, would be valuable for comprehending the full potential and limitations of the model.

References

Chen, Y., Dai, X., Chen, D., Liu, M., Dong, X., Yuan, L., & Liu, Z. (2022). Mobile-Former: Bridging MobileNet and Transformer. CVPR.

Ding, X., Guo, Y., Ding, G., & Han, J. (2019). ACNet. ICCV.

Ding, X., Zhang, X., Han, J., & Ding, G. (2021a). Diverse Branch Block. CVPR.

Ding, X., Zhang, X., Han, J., & Ding, G. (2021b). RepMLP. arXiv:2105.01883.

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., & Sun, J. (2021c). RepVGG. CVPR.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An Image is Worth 16x16 Words. ICLR.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. CVPR.

Howard, A. G., Zhu, M., Chen, B., et al. (2017). MobileNets. arXiv:1704.04861.

Li, J., Xia, X., Li, W., et al. (2022a). Next-ViT. arXiv:2207.05501.

Li, Y., Hu, J., Wen, Y., et al. (2022d). EfficientFormer. NeurIPS.

Liu, Z., Lin, Y., Cao, Y., et al. (2021). Swin Transformer. ICCV.

Liu, Z., Mao, H., Wu, C. Y., et al. (2022). ConvNeXt. CVPR.

Ma, N., Zhang, X., Zheng, H., & Sun, J. (2018). ShuffleNet V2. ECCV.

Mehta, S., & Rastegari, M. (2022). MobileViT. ICLR.

Sandler, M., Howard, A. G., Zhu, M., et al. (2018). MobileNetV2. CVPR.

Srinivas, A., Lin, T., Parmar, N., et al. (2021). Bottleneck Transformers. CVPR.

Touvron, H., Cord, M., Douze, M., et al. (2021). DeiT. ICML.

Vasu, P. K. A., Gabriel, J., Zhu, J., Tuzel, O., & Ranjan, A. (2022). MobileOne. arXiv:2206.04040.