Frequency Domain Adaptive Filters in Vision Transformers

Oscar Ondeng[1], Heywood Ouma[1], Peter Akuon[1]
[1]University of Nairobi, Department of Electrical and Information Engineering, Nairobi, Kenya

Introduction

1. Problem

Vision Transformers (ViTs) have revolutionized computer vision, achieving state-of-the-art performance in many computer vision tasks. However, they face a number of inherent challenges:

- Self-attention mechanism has a **quadratic computational complexity (O(n²))**
  - This makes training **expensive**
- Lack of spatial inductive biases
- Hence requiring large-scale pre-training for optimal performance

2. Solution

Introduction of a Multi-Head Adaptive Filter Frequency Vision Transformer (MAF-FViT), which **replaces the self-attention mechanism** with **frequency-domain adaptive filters**:

- Leverages multi-head adaptive filtering in the frequency domain
- Reduces computational complexity to log-linear, instead of quadratic

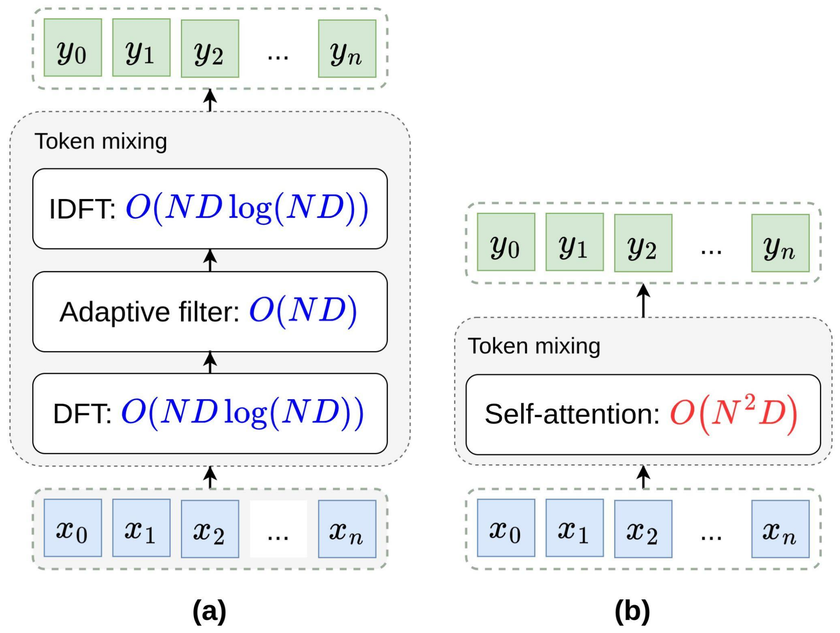

Fig. 1 *The core token mixers of (a) the Multi-head Adaptive Filter - Frequency ViT (MAF-FViT) with log-linear complexity, and (b) the traditional self-attention with quadratic complexity.*

Current Work

1. Objectives

- Develop models that reduce the computational complexity of Transformers
- Develop smarter algorithms that increase effectiveness without relying solely on larger models and more data
- Incorporate mechanisms into ViTs to compensate for their lack of inductive biases and mitigate the need for extensive pre-training.

2. Methodology

- **Core Idea:** Replace the standard self-attention mechanism with multi-head adaptive filtering applied in the frequency domain
- **Datasets:** CIFAR-10, CIFAR-100, and Street View House Numbers (SVHN) => **small-scale datasets** for image classification
- **Metrics:** Top-1 classification accuracy, training & inference time, parameter count
- **Training:** from scratch, on NVIDIA A30 24GB GPU

3. Results

- **Overall Performance**: achieves higher or comparable performance to self-attention
- **Computational Complexity**: **O(ND log (ND))**. Lower FLOPS, parameter count, and faster inference time
- Outperforms other Fourier-Transform-based methods: GFNet, AFNO, and AFFNet

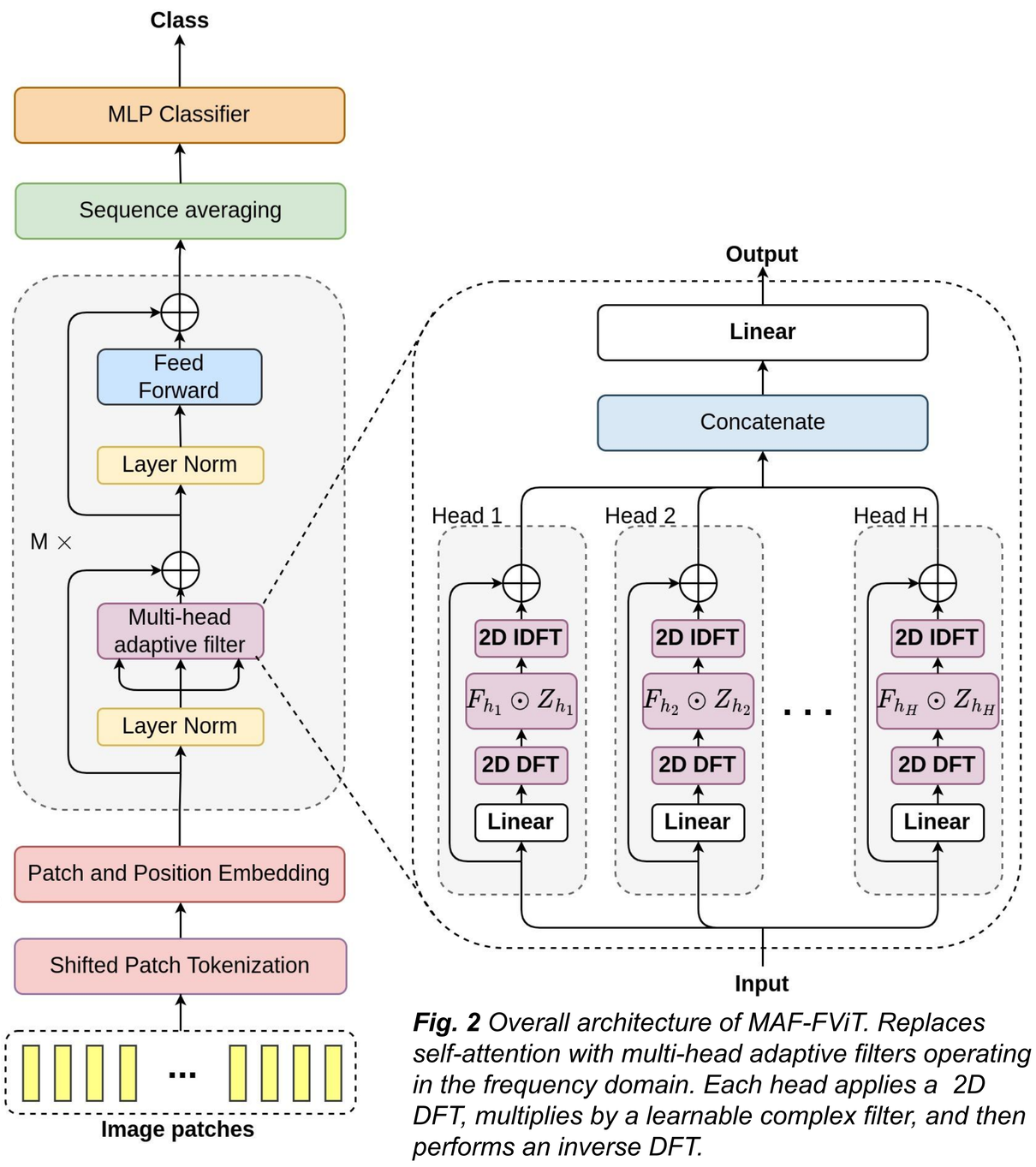

Fig. 2 *Overall architecture of MAF-FViT. Replaces self-attention with multi-head adaptive filters operating in the frequency domain. Each head applies a 2D DFT, multiplies by a learnable complex filter, and then performs an inverse DFT.*

Conclusion & Expectations

- Frequency-domain adaptive filters are a scalable and efficient alternative to self-attention for Vision Transformers, especially for small-scale datasets and resource-constrained environments
- MAF-FViT achieves log-linear complexity, better performance (accuracy, FLOPs, inference/training times), and a lower parameter count compared to other models
- The multi-head configuration of adaptive filters enhances the model's ability to capture complex patterns, optimizing the balance between efficiency and accuracy
- **Future work:** extension to large-scale datasets, exploring other transforms

Acknowledgment

UNIVERSITY OF NAIROBI
kenet Kenya Education Network
AFRICAN DEVELOPMENT BANK GROUP
REPUBLIC OF KENYA Ministry of Education

Contact: oscaror@uonbi.ac.ke