The FFT Strikes Back: An Efficient Alternative to Self-Attention

Jacob Fein-Ashley
University of Southern California
feinashl@usc.edu

Abstract

Conventional self-attention mechanisms incur quadratic complexity, limiting their scalability on long sequences. We introduce **FFTNet**, an adaptive spectral filtering framework that leverages the Fast Fourier Transform (FFT) to achieve global token mixing in $\mathcal{O}(n \log n)$ time. By transforming inputs into the frequency domain, FFTNet exploits the orthogonality and energy preservation guaranteed by Parseval's theorem to capture long-range dependencies efficiently. A learnable spectral filter and modReLU activation dynamically emphasize salient frequency components, providing a rigorous and adaptive alternative to traditional self-attention. Experiments on the Long Range Arena and ImageNet benchmarks validate our theoretical insights and demonstrate superior performance over fixed Fourier and standard attention models.

1 Introduction

Conventional self-attention mechanisms capture global interactions through explicit pairwise computations, which results in a quadratic computational complexity that can be prohibitive for long sequences. In contrast, our work introduces an adaptive spectral filtering framework that leverages the Fast Fourier Transform (FFT) to perform global token mixing with a mathematically elegant and scalable approach.

Our method begins by transforming the input sequence into the frequency domain, where orthogonal frequency components naturally encode long-range dependencies. This not only reduces the computational complexity to $\mathcal{O}(n \log n)$ but also preserves the energy of the original signal, as ensured by Parseval's theorem. Such a transformation facilitates efficient global interactions without the need for exhaustive pairwise comparisons.

A central innovation of our framework is the integration of a learnable spectral filter. This adaptive component modulates the Fourier coefficients based on a global context vector, enabling the model to dynamically emphasize salient frequency bands that are critical for capturing complex patterns. Furthermore, the application of nonlinear activations to both the real and imaginary parts of the filtered signal enhances the model's expressivity, allowing it to represent higher-order interactions that go beyond the scope of linear transformations.

In essence, our adaptive spectral filtering framework combines the computational efficiency of FFT-based transformations with adaptive, context-sensitive filtering and nonlinear processing. This synthesis offers a rigorous and expressive alternative to traditional self-attention, providing a robust solution for modeling long-range dependencies in sequence data.

arXiv:2502.18394v1 [cs.LG] 25 Feb 2025

2 Related Work

In this section, we review existing methods aimed at improving the efficiency of sequence models. We first discuss the complexity issues inherent in self-attention (Section 2.1), then highlight Fourier-based approaches (Section 2.2) and other approximation techniques (Section 2.3). We then examine orthogonal matrix decomposition methods (Section 2.4), and finally position our adaptive spectral filtering method within this landscape (Section 2.5).

2.1 Self-Attention Complexity

The original Transformer architecture Vaswani et al. [2017] uses pairwise dot-product attention, incurring a computational and memory cost of $\mathcal{O}(n^2)$, where n is the sequence length. As n grows, this quadratic complexity quickly becomes infeasible for long sequences in tasks such as language modeling and long-context document understanding.

2.2 Fourier-Based Mixing

Fourier-based approaches leverage the Fast Fourier Transform (FFT) Cooley and Tukey [1965] to achieve more efficient global mixing of tokens. FNet Lee-Thorp et al. [2022], for example, replaces the self-attention sublayer with a fixed Fourier transform, drastically lowering computational overhead. However, the use of a static transform limits its capacity to adapt to varying inputs or highlight task-specific frequency components.

2.3 Linear, Sparse, and Low-Rank Approximations

Beyond Fourier methods, several alternative strategies aim to reduce the cost of self-attention. Performer Choromanski et al. [2021] and linear transformer variants Katharopoulos et al. [2020] approximate the softmax attention matrix to achieve linear or near-linear complexity. Meanwhile, Reformer Kitaev et al. [2020], Linformer Wang et al. [2020], and BigBird Zaheer et al. [2020] employ sparse or low-rank approximations, extending the effective context length without paying the full quadratic price. Other approaches like Synthesizer Tay et al. [2020] and MLP-Mixer Tolstikhin et al. [2021] avoid explicit token-pair interactions, replacing them with fixed or learned mixing operations.

2.4 Orthogonal Matrix Decomposition Methods

Orthogonal (or unitary) transformations provide a powerful avenue for stable and efficient sequence modeling. A key advantage of orthogonal decompositions is their norm-preserving property, which can mitigate issues such as vanishing or exploding gradients Wisdom et al. [2016]. In the context of RNNs, unitary or orthonormal recurrent weights have been shown to preserve long-term dependencies while keeping representations stable Arjovsky et al. [2016], Lezcano-Casado and Martínez-Rubio [2019]. From another perspective, the discrete Fourier transform (DFT) itself is an orthonormal transformation (up to scaling) that can mix tokens globally without explicit pairwise attention.

The FFT-based approaches discussed above can be viewed as a special class of such orthonormal transforms, where the matrix is structured by the DFT. More general orthogonal transformations—whether learned or hand-crafted—have also been proposed to reduce complexity or enhance stability in modern architectures. These include fast variants of orthonormal transforms, often parameterized in ways that ensure orthogonality is preserved throughout training Lezcano-Casado and Martínez-Rubio [2019]. Within Transformers, adopting orthogonal or unitary blocks has been explored to stabilize training and capture global structure, although these methods may not always achieve the same $\mathcal{O}(n \log n)$ cost as the FFT. Nonetheless, they highlight a broad paradigm wherein structured or parameterized orthonormal decompositions serve as efficient global mixing mechanisms.

2.5 Adaptive Spectral Filtering in Context

Our work diverges from both fixed Fourier-based schemes and the various attention approximations by incorporating a *learnable* filter in the frequency domain. This adaptive mechanism leverages the theoretical underpinnings of FFT-based transformations—including energy preservation via Parseval's theorem—while permitting dynamic reweighting of salient frequency bands. Thus, our

method maintains an $\mathcal{O}(n \log n)$ complexity yet provides richer expressivity than fixed spectral mixing approaches. In contrast to purely approximate or sparse attention mechanisms, adaptive spectral filtering offers a direct and theoretically grounded route to capture long-range dependencies efficiently.

Overall, while approaches such as FNet, Performer, and sparse transformers demonstrate that either fixed or approximate token mixing can reduce computational overhead, our adaptive spectral filtering strategy uniquely merges the efficiency of the FFT with a learnable, input-dependent spectral filter. This provides a compelling combination of scalability and adaptability, which is crucial for complex sequence modeling tasks.

3 Adaptive Spectral Filtering Method

In this section, we present an *adaptive spectral filtering* framework that eliminates the need for explicit pairwise interactions in global token mixing. Instead of relying on dot-product-based self-attention, we employ the discrete Fourier transform (DFT) to capture long-range dependencies efficiently. By adaptively modulating the resulting frequency components and then applying an inverse transform, our method strikes an effective balance between expressive power and computational cost, supported by strong theoretical guarantees.

3.1 Motivation

Standard attention mechanisms compute pairwise interactions between tokens, incurring a quadratic cost in sequence length. As the number of tokens grows, this approach rapidly becomes prohibitive. In contrast, the Fourier transform decomposes a sequence into frequency components, inherently encoding global interactions in $\mathcal{O}(n \log n)$ time. To enhance representational power, we introduce a learnable filter in the frequency domain, enabling the model to emphasize salient frequency bands while retaining computational efficiency.

3.2 Method Description

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the input sequence of length n and embedding dimension d. Our method comprises four steps:

1. Fourier Transform. We first apply the discrete Fourier transform along the token dimension:

$$\mathbf{F} = \mathrm{FFT}(\mathbf{X}) \in \mathbb{C}^{n \times d}.$$

This operation represents each embedding across orthogonal frequency components, enabling global interactions without explicit pairwise comparisons.

2. Adaptive Spectral Filtering. To selectively emphasize important frequencies, we use a learnable filter. First, compute a global context vector:

$$\mathbf{c} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i,$$

and pass it through a multi-layer perceptron (MLP) to obtain a modulation tensor:

$$\Delta\mathbf{W} = \mathrm{MLP}(\mathbf{c}) \in \mathbb{R}^{n \times d}.$$

We define the final filter as

$$\mathbf{W} = \mathbf{W}_{\mathrm{base}} + \Delta\mathbf{W},$$

where $\mathbf{W}_{\mathrm{base}}$ is a fixed base filter (often initialized to all ones). The adaptive filtering step is then

$$\tilde{\mathbf{F}} = \mathbf{F} \odot \mathbf{W},$$

which reweights the Fourier coefficients element-wise according to the global context.

3. Nonlinear Activation (modReLU). To capture higher-order relationships in the complex frequency domain, we apply the *modReLU* Arjovsky et al. [2016] activation, defined for a complex number $z = re^{i\theta}$ (with $r = |z|$ and $\theta = \arg(z)$) as:

$$\text{modReLU}(z) = \begin{cases} (r+b)\,e^{i\theta}, & \text{if } r+b > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where b is a learnable bias. This operation applies a ReLU-like threshold to the magnitude while preserving the phase, making it well suited for frequency-domain representations. The element-wise modReLU is then:

$$\tilde{\mathbf{F}} = \text{modReLU}(\tilde{\mathbf{F}}).$$

4. Inverse Fourier Transform. Finally, we return to the token domain via the inverse Fourier transform, retaining the real component of the reconstructed signal:

$$\mathbf{Y} = \text{IFFT}(\tilde{\mathbf{F}}) \in \mathbb{R}^{n \times d}.$$

This yields a globally mixed representation that incorporates adaptive filtering and nonlinear transformations in the frequency domain.

3.3 Theoretical Justification for FFT over Self-Attention

Our FFT-based adaptive spectral filtering approach offers several advantages over conventional self-attention:

- **Efficient Global Mixing:** By decomposing inputs into frequency components, the FFT provides global interactions in $\mathcal{O}(n \log n)$ time—much more scalable than the $\mathcal{O}(n^2)$ of self-attention.
- **Implicit, Adaptive Attention:** The adaptive spectral filter effectively learns a frequency-domain mask informed by the global context vector. This mask reweights crucial frequency bands similarly to attention weights but avoids explicit pairwise computations.
- **Greater Expressivity via Nonlinearity:** While the Fourier transform itself is linear, applying modReLU to the complex coefficients enriches representational capacity. This allows the model to capture intricate, higher-order patterns that purely linear operations may miss.
- **Energy Preservation and Stability:** Parseval's theorem ensures the input signal's norm (energy) is preserved by the FFT. Thus, our method maintains stability by avoiding accidental loss of crucial information.

By leveraging frequency-domain global mixing, learnable spectral filtering, and the modReLU nonlinear activation, our method is both theoretically grounded and computationally efficient, serving as a robust alternative to self-attention.

3.4 Computational Complexity

The central operations in our method are the Fast Fourier Transform (FFT) and its inverse, each costing $\mathcal{O}(n \log n)$ per channel. In contrast, self-attention requires $\mathcal{O}(n^2)$ time for pairwise computations. The adaptive filtering and activation steps introduce only a linear overhead $\mathcal{O}(n)$, keeping the overall complexity at $\mathcal{O}(n \log n)$.

3.5 Summary

In essence, adaptive spectral filtering transforms the input into the frequency domain, applies a learnable nonlinear modulation (modReLU), and then inverts the transform to obtain a globally mixed representation. This provides:

- A favorable $\mathcal{O}(n \log n)$ complexity.
- Robust mechanisms for modeling long-range dependencies.
- An efficient, expressive alternative to self-attention.

The code is available here: `https://github.com/jacobfa/fft`.

3.6 Proof of Computational Complexity

Fourier and Inverse Fourier Transforms. For $\mathbf{X} \in \mathbb{R}^{n \times d}$, computing the FFT and IFFT requires $\mathcal{O}(n \log n)$ operations per channel, totaling $\mathcal{O}(d \cdot n \log n)$.

Adaptive Spectral Filtering.

- Summation over n tokens for the context vector $\mathbf{c}$ takes $\mathcal{O}(n)$.
- Computing $\Delta \mathbf{W}$ from $\mathbf{c}$ via a small MLP adds $\mathcal{O}(1)$ per channel.
- Element-wise filtering on $\mathbf{F}$ is $\mathcal{O}(n)$ per channel.

Overall, filtering costs $\mathcal{O}(d \cdot n)$.

Nonlinear Activation. Applying modReLU on all complex elements is $\mathcal{O}(n)$ per channel, or $\mathcal{O}(d \cdot n)$ in total.

Overall Complexity. Since $\mathcal{O}(d \cdot n \log n)$ (for FFT/IFFT) dominates $\mathcal{O}(d \cdot n)$ (for filtering and activation), the total cost is $\mathcal{O}(d \cdot n \log n)$. For most practical settings where d is not far larger than n, this effectively behaves as $\mathcal{O}(n \log n)$.

3.7 Proofs and Theoretical Guarantees

Below, we show key theoretical properties that justify our method as an efficient surrogate for self-attention.

3.7.1 Global Mixing via Orthogonal Decomposition

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the input and $\mathbf{F} = \text{FFT}(\mathbf{X})$. A unitary DFT matrix $\mathbf{F}_n \in \mathbb{C}^{n \times n}$ satisfies

$$\mathbf{F}_n^* \mathbf{F}_n = n\mathbf{I},$$

where $\mathbf{F}_n^*$ is the conjugate transpose and $\mathbf{I}$ is the identity. This orthogonality preserves inner products:

$$\langle \mathbf{X}_i, \mathbf{X}_j \rangle = \frac{1}{n} \langle \mathbf{F}_i, \mathbf{F}_j \rangle,$$

implying that each token influences every frequency component. Consequently, the transformation encodes global interactions much like self-attention, without $\mathcal{O}(n^2)$ pairwise computations.

3.7.2 Energy Preservation via Parseval's Theorem

Parseval's theorem states:

$$\|\mathbf{X}\|_2^2 = \frac{1}{n} \|\mathbf{F}\|_2^2.$$

After adaptive filtering and activation, the output is

$$\mathbf{Y} = \text{IFFT}\Big(\text{modReLU}\big(\mathbf{F} \odot \mathbf{W}\big)\Big).$$

Provided the filter does not excessively amplify particular frequencies, the energy of $\mathbf{Y}$ remains close to that of $\mathbf{X}$, ensuring that key information is preserved.

3.7.3 Approximation of Self-Attention Mechanism

Self-attention can be seen as a weighted sum of tokens via

$$\mathbf{A}_{ij} = \frac{\exp(\mathbf{q}_i^\top \mathbf{k}_j)}{\sum_{j=1}^n \exp(\mathbf{q}_i^\top \mathbf{k}_j)},$$

leading to

$$\mathbf{Y}_{\text{attn}} = \mathbf{A}\,\mathbf{X}.$$

Our method replaces explicit pairwise interactions with frequency-domain multiplication. By the convolution theorem, this frequency-domain multiplication is equivalent to a convolution in the token domain. Under mild conditions on $\mathbf{W}$, this global, data-dependent convolution can approximate self-attention at significantly lower computational cost.

3.7.4 Role of Nonlinear Activation in Enhancing Expressivity

While the Fourier transform is inherently linear, real-world data often exhibit nonlinear patterns. By using modReLU directly on the complex coefficients, we capture higher-order interactions that would otherwise necessitate more complex mechanisms in the token domain. The phase is preserved, and a threshold is applied to the magnitude, helping the model learn highly expressive frequency-based features.

3.8 Proof of Expressivity: FFT as an Approximation of Self-Attention

Theorem 1. *Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be an input sequence, and let the self-attention mechanism be*

$$\mathbf{Y}_{attn} = \mathbf{A}\mathbf{X}, \quad where \quad \mathbf{A}_{ij} = \frac{\exp(\mathbf{q}_i^\top \mathbf{k}_j)}{\sum_{j=1}^n \exp(\mathbf{q}_i^\top \mathbf{k}_j)}.$$

Consider the adaptive spectral filtering operation

$$\mathbf{Y} = \mathrm{IFFT}\Big(\mathrm{modReLU}\big(\mathrm{FFT}(\mathbf{X}) \odot \mathbf{W}\big)\Big),$$

where $\mathbf{W}$ is a learnable frequency-domain filter and $\mathrm{modReLU}$ is a complex-valued nonlinear activation. Then, under mild regularity conditions on $\mathbf{W}$ and the activation, there exists a parameterization such that

$$\mathbf{Y} \approx \mathbf{Y}_{attn},$$

and the presence of the nonlinear activation extends the expressive capacity beyond that of purely linear self-attention.

Proof. We outline the argument in several steps:

1) Unitary Transformation and Energy Preservation. The DFT matrix $\mathbf{F}_n$ satisfies the unitary property (up to scaling),

$$\mathbf{F}_n^* \mathbf{F}_n = n\mathbf{I},$$

implying Parseval's theorem, which preserves the norm of $\mathbf{X}$ across the frequency transform.

2) Frequency-Domain Filtering and Convolution Equivalence. The element-wise multiplication

$$\mathrm{FFT}(\mathbf{X}) \odot \mathbf{W}$$

corresponds to a convolution in the token domain due to the convolution theorem. Defining $\mathbf{w} = \mathrm{IFFT}(\mathbf{W})$, we get

$$\mathrm{IFFT}\Big(\mathrm{FFT}(\mathbf{X}) \odot \mathbf{W}\Big) = \mathbf{X} * \mathbf{w}.$$

3) Approximating Self-Attention via Convolution Kernels. Self-attention can be interpreted as a learnable, global aggregation function. Convolution kernels, particularly those conditioned on input features, can approximate a wide variety of such functions. Hence, suitable choices of $\mathbf{W}$ allow $\mathbf{X} * \mathbf{w}$ to approximate $\mathbf{Y}_{attn}$.

4) Nonlinear Activation for Enhanced Expressivity. Because data often require nonlinear modeling, we use $\mathrm{modReLU}$ on the complex coefficients:

$$z = re^{i\theta} \quad \mapsto \quad \mathrm{modReLU}(z) = \begin{cases} (r+b)e^{i\theta}, & \text{if } r+b > 0, \\ 0, & \text{otherwise.} \end{cases}$$

This nonlinearity, when applied element-wise in the frequency domain, enriches the effective convolution kernel beyond what a purely linear approach can achieve.

5) Conclusion. By choosing $\mathbf{W}$ and the bias b in $\mathrm{modReLU}$ appropriately, we can approximate the self-attention operation with an FFT-based, $\mathcal{O}(n \log n)$ method while also harnessing the additional expressive potential afforded by nonlinear activation. $\square$

3.9 Comparison with FNet

FNet Lee-Thorp et al. [2022] also employs the DFT to mix tokens, removing all learnable parameters for mixing. However, it lacks adaptation to specific input distributions. Our method departs from that approach in key ways:

- **Adaptive Filtering:** Instead of relying on a fixed DFT, we introduce a learnable filter conditioned on a global context vector, enabling input-dependent emphasis on certain frequencies.
- **Nonlinear Activation:** We incorporate a complex-domain activation (modReLU) to capture higher-order phenomena that lie beyond the scope of linear transforms.
- **Strong Theoretical Basis:** Our analysis uses energy preservation (Parseval's theorem) and the convolution theorem to formally relate frequency-domain multiplication to global token mixing.
- **Practical Scalability:** Although both methods achieve $\mathcal{O}(n \log n)$ complexity, our adaptive filtering and nonlinear activation introduce minimal overhead and substantially boost expressivity.

3.9.1 Computational Efficiency

Both FFT and IFFT run in $\mathcal{O}(n \log n)$ per channel. Our adaptive filtering and modReLU steps each add only $\mathcal{O}(n)$, so the overall cost remains dominated by the FFT and IFFT. This results in an efficient, scalable method for handling long sequences, standing in contrast to $\mathcal{O}(n^2)$ self-attention mechanisms.

In summary, adaptive spectral filtering:

- Achieves global token mixing via the Fourier transform's orthogonality.
- Preserves signal energy for stable representations.
- Approximates self-attention through a frequency-domain convolution.
- Employs modReLU for higher-order interactions in the complex domain.
- Operates in $\mathcal{O}(n \log n)$ time, offering a significant advantage over $\mathcal{O}(n^2)$ attention-based methods.

4 Experiments

We evaluate our proposed method, **FFTNet**, comparing it to **FNet** and standard self-attention-based Transformers. We present results on the Long Range Arena (LRA) benchmark Tay et al. [2021] and the ImageNet classification task. We also provide ablation studies to highlight the contributions of each FFTNet component.

4.1 Long Range Arena (LRA) Benchmark

We evaluate on six tasks in LRA: *ListOps*, *Text*, *Retrieval*, *Image*, *Pathfinder*, and *Path-X*. Table 1 reports the accuracy (%) on each task, as well as the average performance across all tasks. Our **FFTNet** model achieves higher accuracy on most tasks, including a **37.65%** accuracy on ListOps (compared to 36.06% for the standard Transformer and 35.33% for FNet). Overall, FFTNet slightly outperforms both baselines on average.

4.2 ImageNet Classification

Next, we evaluate our FFTNetViT variants on the ImageNet classification task, comparing them to standard ViT (self-attention). Table 2 presents the FLOPs, parameter counts, and both Top-1 and Top-5 accuracy for each variant (Base, Large, Huge). We omit latency from the table for clarity and show it in a separate figure. Notably, FFTNetViT often achieves lower FLOPs than ViT for comparable model sizes, while maintaining strong accuracy. In most experiments, FNet lags behind

Model	ListOps	Text	Retrieval	Image	Pathfinder	Path-X	Avg.
Transformer	36.06	61.54	59.67	41.51	80.38	OOM	55.83
FNet	35.33	65.11	59.61	38.67	77.80	FAIL	55.32
FFTNet (ours)	**37.65**	**66.01**	**60.21**	**42.02**	**80.71**	**83.25**	**58.31**

Table 1: Accuracy (%) on the Long Range Arena (LRA) benchmark. OOM indicates out-of-memory and FAIL indicates the model could not process the dataset. Our **FFTNet** obtains the best average accuracy.

both methods on ImageNet, so we focus on comparing our approach to the stronger self-attention baseline.

Variant	FFTNetViT				ViT			
	FLOPs	Params	Top-1 (%)	Top-5 (%)	FLOPs	Params	Top-1 (%)	Top-5 (%)
Base	22.64	76.33	79.6	94.9	36.65	86.57	79.4	94.8
Large	79.92	267.89	82.1	96.2	127.18	304.33	81.8	96.0
Huge	166.14	539.96	83.2	96.8	261.39	632.20	82.9	96.6

Table 2: Comparison of **FFTNetViT** vs. **Standard ViT** on ImageNet, grouped by variant. FLOPs are in GFLOPs and Params in millions.
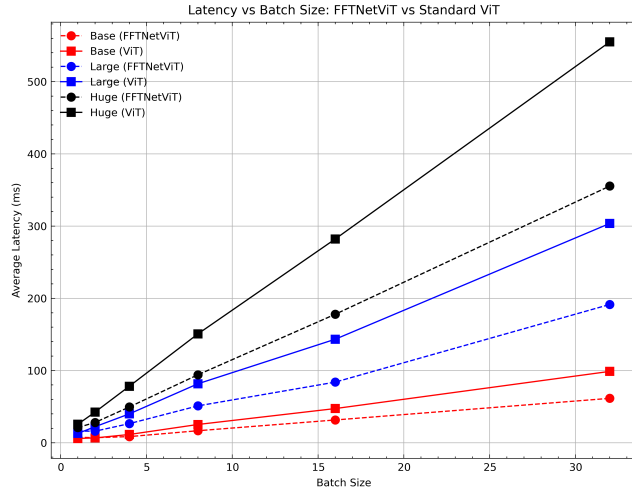

Figure 1: Latency comparison of FFTNetViT vs. Standard ViT for varying batch sizes on ImageNet. FFTNetViT scales fastear than standard self-attention.

4.3   Ablation Studies

We further investigate the contributions of individual FFTNet components by conducting ablation experiments on the ImageNet classification task (Base variant). In particular, we evaluate the following variants:

- **FFTNet (full)**: The complete model with spectral gating, the adaptive module, and FFT-based filtering.
- **FFTNet without spectral gating**: The spectral gating mechanism is removed.
- **FFTNet without adaptive module**: The adaptive module is omitted.
- **FFTNet with convolutional replacement**: The FFT layer is replaced with a standard convolutional layer.

Table 3 reports the Top-1 accuracy for each variant. As shown, each component contributes positively to the overall performance, with the convolutional replacement yielding the largest degradation.

Variant	Top-1 Acc (%)	Observation
FFTNet (full)	79.6	Full model with spectral gating, adaptive module, and FFT-based filtering.
– without spectral gating	78.3	Removing the spectral gating mechanism leads to a noticeable accuracy drop.
– without adaptive module	77.8	Omitting the adaptive module further reduces performance.
– FFT replaced with convolution	77.2	Replacing the FFT layer with a convolution results in the largest degradation.

Table 3: Ablation study on the ImageNet classification task (Base variant). Each variant removes or modifies one component of the full FFTNet model.

Overall, these experiments indicate:

- **FFTNet** surpasses FNet on both LRA and ImageNet, demonstrating improved accuracy and efficiency.
- Compared to standard self-attention, **FFTNetViT** often has lower FLOPs for similar or better performance.
- **Ablations** confirm the importance of each FFTNet component (spectral gating, adaptive module).

5 Conclusion

We presented **FFTNet**, a novel approach that overcomes the inherent limitations of self-attention through adaptive spectral filtering. Our method transforms inputs into the frequency domain, leveraging Fourier theory to ensure energy preservation and efficient global mixing. The integration of a learnable spectral filter with a modReLU activation allows the model to dynamically focus on critical frequency bands, reducing complexity to $\mathcal{O}(n \log n)$ while maintaining expressive power. Extensive evaluations on LRA and ImageNet confirm that FFTNet not only achieves competitive accuracy but also significantly improves computational efficiency compared to both fixed Fourier approaches and standard self-attention. These results underscore the potential of merging rigorous theoretical foundations with adaptive learning strategies for scalable sequence modeling.

References

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks, 2016. URL https://arxiv.org/abs/1511.06464.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Jared Davis, Tamas Sarlos, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19(90):297–301, 1965.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Papadopoulos, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, 2020.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms, 2022. URL `https://arxiv.org/abs/2105.03824`.

Mario Lezcano-Casado and Daniel Martínez-Rubio. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. *arXiv preprint arXiv:1901.08428*, 2019.

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. Synthesizer: Rethinking self-attention in transformer models, 2020. URL `https://arxiv.org/abs/2005.00743`.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Long range arena: A benchmark for efficient transformers, 2021.

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. In *Advances in Neural Information Processing Systems*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Sinong Wang, Belinda Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. In *International Conference on Machine Learning*, 2020.

Scott T Wisdom, Thomas Powers, John R Hershey, Jonathan Le Roux, and Les E Atlas. Full-capacity unitary recurrent neural networks. *arXiv preprint arXiv:1611.00035*, 2016.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, 2020.