Re-implementation of Image Classification using Global Filter Networks

Anonymous submission

Abstract

This project investigates the use of Global Filter Networks (GFNet) as an efficient and effective alternative to traditional models in image classification tasks. While Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) have become standard in computer vision, they often suffer from high computational complexity, especially when capturing long-range dependencies across image pixels. GFNet addresses these limitations by applying global filtering in the frequency domain, enabling a log-linear computational complexity that reduces resource demands without sacrificing accuracy. GFNet was compared to self-attention models on CIFAR10 and CIFAR100, evaluating their performance on accuracy and training efficiency. Results indicate that GFNet with global filters achieved marked improvements in accuracy, increasing by 23% on CIFAR10 and 14% on CIFAR100, with training times reduced by approximately 25–27% compared to baseline self-attention models. However, the observed filter patterns suggest that the model could benefit from further optimization to fully capture frequency-domain features. These findings highlight GFNet's potential for scalable, accurate image classification, helping with future work on optimization, larger datasets, and real-world applications.

Code — https://anonymous.4open.science/r/ECE570-code-submission-6F8A/

Introduction

Image classification is a foundational task in computer vision, focused on assigning labels to images based on their content. This task is critical in numerous applications, such as medical diagnostics, where images are analyzed to identify potential health issues, and social media, where images are categorized and tagged for efficient content management and retrieval. As such, advancements in image classification models are central to the progression of AI technologies, serving as a building block for more complex tasks like object detection, image segmentation, and action recognition in video analysis.

Traditionally, Convolutional Neural Networks (CNNs) have been the go-to model for image classification, owing to their ability to extract hierarchical features and learn both local and global patterns. More recently, Vision Transformers (ViTs) have emerged, leveraging self-attention mechanisms to capture long-range dependencies across pixels, enabling the model to process complex spatial relationships. However, a key limitation of these approaches lies in their computational complexity, particularly for high-resolution images. ViTs, in particular, suffer from quadratic complexity with respect to the number of patches, making them less scalable and computationally expensive for large-scale applications.

This project investigates Global Filter Networks (GFNet), a model architecture that seeks to address these limitations by processing images in the frequency domain. GFNet applies discrete Fourier transforms to convert spatial image data into the frequency domain, where it performs global filtering to capture both local and global dependencies efficiently. This unique approach reduces the model's computational complexity to log-linear, offering a promising solution for scalable and efficient image classification without the high computational demands associated with self-attention in transformers or the depth required in CNNs. By focusing on global frequency features, GFNet can potentially achieve similar or improved classification performance at a fraction of the computational cost.

The main objective of this project is to evaluate GFNet's performance in image classification tasks on benchmark datasets, specifically CIFAR10 and CIFAR100. This study aims to assess whether GFNet can balance classification accuracy and computational efficiency, comparing its effectiveness to that of CNNs and ViTs. Key metrics such as accuracy, training speed, and computational complexity will be used to measure GFNet's viability as an alternative to existing models. Understanding how GFNet performs in these aspects is essential for determining its potential use in real-world applications where both efficiency and accuracy are crucial.

Our approach involves implementing and training a GFNet model on CIFAR10 and CIFAR100, analyzing its ability to classify images accurately and efficiently. By comparing GFNet to conventional models, we aim to explore the practical advantages of frequency-domain processing for image classification, highlighting GFNet's scalability and applicability in resource-constrained environments. This research is particularly relevant to ongoing efforts to develop machine learning models that are not only high-performing but also computationally feasible for large-scale deployment, advancing the field toward more accessible and efficient AI solutions.

Related Work

The course project was based on work from another paper (Rao et al. 2021). The project was planned to be a re-implementation using global filters, however, there are still many differences between the original code and the code implemented in this project.

Global Filter Networks for Image Classification

This paper presented the GFNet and included the methodology and experiments used. They explain that GFNet intends to achieve competitive performance by replacing the heavy quadratic complexity self-attention layer with a simpler and more efficient one. This would include having the global filter layer with the fourier transform and the global filters, followed by a feed forward network that would include the multi-layer perceptron model (MLP).

The experiments focus on evaluating GFNet's performance on image classification and transfer learning tasks compared to many of the other models, tested on a variety of datasets, including ImageNet, CIFAR-10/100, Stanford Cars and Flowers-102.

Upon testing on the ImageNet dataset, GFNet showed improvements over recent MLP-based models and competitive performance against vision transformers. Multiple GFNet variants (e.g., GFNet-Ti, GFNet-S) are benchmarked, demonstrating superior trade-offs of accuracy for complexity.

GFNet was also evaluated on transfer learning benchmarks such as CIFAR-10, CIFAR-100, Stanford Cars, and Flowers-102. They found that GFNet performed well, consistently outperforming ResMLP models and achieving comparable results to EfficientNet.

GFNet showed significant improvements in memory usage and latency over transformers and MLP models when processing high-resolution images, with this scalability being attributed to the log-linear complexity of the global filter layer, as predicted.

GFNet was also tested for robustness using adversarial attacks (FGSM, PGD) and out-of-distribution datasets (ImageNet-A, ImageNet-C). It demonstrated favorable robustness and generalization ability compared to baseline models.

The experiments highlighted the efficiency, robustness, and generalization potential of GFNet in various image classification and transfer learning tasks. The paper itself provides a great introduction to GFNets and has all the requisite experiments, however, there is still further that can be explored with the topic. There is a lack of direct experiments, meaning that the paper demonstrated the raw capabilities of GFNet, while limiting the scope of the conclusions somewhat.

Fast Fourier Convolution

The goal of this paper was the exposition of unit called the fast Fourier convolution (FFC) and fast Fourier transform (FFT) (Chi, Jiang, and Mu 2020). This fast Fourier transform was a part of the process in GFNets, among many other models, including not just image classification, but also video action recognition and human keypoint detection. FFC uses the FFT to capture global information in the frequency domain and combines it with local convolution operations. This creates a convolutional unit with both local and global receptive fields, facilitating information flow across different scales within a single unit.

The architecture splits the input into local and global paths. The local path performs standard convolutions on a portion of the input channels, while the global path processes the remaining channels using Fourier transforms. After processing, the outputs from both paths are aggregated. This makes FFC adaptable and suitable for replacing traditional convolutions in existing architectures without major adjustments.

The paper experimented with ImageNet for image recognition, Kinetics for video action recognition, and MSCOCO for human keypoint detection, and found consistent elevations in accuracy by significant margins for each experiment. With image classification in particular, the researchers found that FFC-ResNet-50 (a pre-existing model ResNet-50 with FFC substituted in) had a 0.4% better accuracy than the base model while costing only 60% of the parameters. They found it was easy to substitute and was effective in multiple models.

The paper focused on some of the simpler tasks, which included image classification, however, there wasn't as much exploration of more complex, denser prediction tasks. In addition, there was limited analysis in cross-scale fusion, which was mentioned as a key benefit to FFC.

Differences from original work

There are some key differences between the implemented project and the original model provided (Rao et al. 2021).

In the reimplementation, the model architecture is kept the same, with all the same layers: patch embedding, global filter, feed forward network with layer norm and MLP, and final linear layer. A similar visualization for the global filters was created as well.

The biggest difference is a great simplification in scope of the project. In the original paper, the authors created multiple unique GFNet models, with two different architectures, using classes called GFNet and GFNetPyramid. The difference between these is that GFNet adopts a transformer-style model that has a fixed number of tokens in each block, while GFNetPyramid adopts a CNN-styled hierarchical model with gradually down sampled tokens. Models were made of each of these classes with varying parameters and each tested. The implementation in this current paper only implements the former architecture, the transformer-style model, and tests with one model using its own parameters, although other models with differing parameters can be easily initialized.

Model testing was a large part of this difference in scope, as well. Some parameters may have differed in the model, along with some difference in parameters during training. A notable example would be the number of epochs used, where the original implementation uses 30, and the reimplementation only used 10. Additionally, in the original paper, many more tests were done with baseline models as well, and tests were also done on many other datasets compared to the new

implementation. The original paper performed its main tests on the ImageNet dataset, along with Flowers-102, Cars-196, CIFAR10, and CIFAR100. The models in this paper were only evaluated with CIFAR10, and CIFAR100, which has some impact on the findings of the experiment. In particular, the number of tokens was unable to be evaluated given that the CIFAR datasets are both 32x32, so there were no results on the scalability of input size. In addition, some of the metrics evaluated in the original paper were omitted, such as parameters, latency, and GPU memory. Robustness and generalization ability were also not evaluated.

Other differences include some minor omittances of customization that the original provided, such as allowing for changing of dropout rates in class initialization, and function used for the activation layer. These parameters can still be changed in the actual class, but have not been implemented as being able to be set during initialization.

Problem Definition

In this project, we address the problem of efficient and accurate image classification, with a focus on evaluating the effectiveness of Global Filter Networks (GFNet) as an alternative to commonly used deep learning models in computer vision. Standard models such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) often rely on extensive layers or attention mechanisms to capture both local and global dependencies, which can lead to high computational costs and inefficiencies, especially as image resolutions increase. This complexity poses limitations in terms of scalability and deployability, particularly for resource-constrained applications.

The core of this project is whether or not GFNet can achieve competitive accuracy in image classification while reducing computational complexity and improving scalability relative to standard vision models. Therefore, this project aims to evaluate GFNet on model complexity and accuracy and help to determine if GFNet is a viable solution in the world of image classification.

Methodology

The implementation of the model required multiple parts, including the model itself, the environment for using the model, decisions on parameter values, training, and evaluating the model.

Global Filter Algorithm

The proposed global filter layer in GFNet serves as an efficient alternative to the self-attention layer by mixing tokens across different spatial locations in the frequency domain. This process begins by applying a 2D Fast Fourier Transform (FFT) to the input tokens, transforming them into the frequency domain. The resulting complex tensor, is then put through element-wise multiplication with the global filter. This filter functions as a set of frequency-specific, learnable filters that modify the spectral information. Afterwards, an inverse FFT transforms the filtered result back to the spatial domain. Additionally, the global filter layer learns distinct patterns in the frequency domain, enabling it to capture relationships more effectively than spatial convolutions.

Model Architecture

The approach was based on the provided model architecture, which included the following components:

- Patch Embedding Layer: The input image is divided into non-overlapping patches using a convolutional layer with a stride and kernel size equal to the patch size. This layer outputs a series of patch embeddings, each of which represents a localized region of the image.
- Global Filter Layer: After patch embedding, we apply a global filter in the frequency domain. This layer performs a Fourier Transform on the patch embeddings to capture long-range dependencies. By multiplying the transformed embeddings with learnable filters in the Fourier domain, the model can enhance or suppress certain frequencies, then use an inverse Fourier Transform to return the modified data to the spatial domain. This operation is parameterized with complex weights initialized for each layer.
- Feed Forward Network (FFN): The output from the global filter layer is passed through a feed-forward network consisting of a layer norm, and an MLP layer, which includes linear layers, a GELU (Gaussian Error Linear Unit) activation function, along with dropout to prevent overfitting. The global filter layer is applied on each iteration as well. This feed-forward network is applied multiple times.
- Finishing Layers: A final dropout and linear layer are applied to the output of the FFN. A final group average pooling layer is omitted to match to the provided model.

In addition, in order to provide a baseline model for comparison, we replaced the global filter blocks with a self-attention block, which is common in many vision transformer models. This replacement allowed us to gauge the actual effectiveness of the global filter. All other layers remained the same, while the global filter was replaced with the following:

- Self-Attention Block: To compare GFNet's performance with other techniques, we experimented with a multi-head self-attention mechanism replacing the global filter layer. This attention mechanism calculates dependencies between patches, allowing each patch to attend to all other patches in the input image.

Experimental Setup

- Environment: The model used a NVIDIA T4 GPU from Google Colab for training and testing, and used PyTorch as the framework.
- Dataset: The model was trained and evaluated on both CIFAR10 and CIFAR100, consisting of 60000 images (50000 for training and 10000 for testing), with 10, and 100 classes, respectively, with images preprocessed to a resolution of 32x32.

- Training Procedure: We trained GFNet using the Adam optimizer with a learning rate of 0.001. The model was trained for 10 epochs with a batch size of 128.
- Loss Function: Cross-entropy loss was used to optimize the model due to its suitability for multi-class classification tasks. The loss function was computed across each batch during training.

Model Parameters

The following include the parameters for the model class that could be set at initialization, with their default values.

- Number of Classes: The number of classes in the dataset was set at default to be 10 to accomodate the CIFAR10 dataset.
- Input Size: The size of the images in the dataset was set at default to 32 to accomodate both the CIFAR10 and CIFAR100 datasets.
- Channels: The number of channels was set to 3 to accomodate RGB channels.
- Patch Size: The patch size was set to 4 to divide each image into 8x8 patches.
- Embedding Dimension: Each patch was projected into a higher-dimensional space with an embedding dimension of 256. This parameter controls the model's representational capacity.

The following include the parameters that were hard coded with their values in to the model.

- Depth: The depth of the FFN, which determined the number of global filter layers, was chosen to be 12.
- MLP Ratio: A MLP ratio of 2 was used for the linear layers, converting between the embed dimension and double the embed dimension.
- Activation Layer: GELU was chosen as the activation function for use in the MLP.
- Dropout: A dropout rate of 0.1 was applied in the feed-forward network to prevent overfitting.
- Self-Attention Parameters: For comparison, multi-head self-attention used 8 heads, with each head operating over a subset of the embedding dimension to calculate scaled dot-product attention between patches.

Evaluation

The model's performance was evaluated via the comparison between the global filter models and the self-attention models on key metrics such as accuracy and loss. Accuracy, a measure of the percentage of correctly classified images, and loss, which reflects the model's error during training and testing, were monitored across both the CIFAR datasets to capture the model's generalization capability and consistency. These metrics provided a quantitative basis for evaluating the models effectiveness relative to standard self-attention mechanisms.

To gain deeper insights into the model's behavior, additional code was implemented to analyze misclassified images. This approach allowed a viewpoint into where the model struggled, examining the specific classes or types of images that were misclassified. By isolating these instances, we could better understand some of the shortcomings of the model.

Moreover, we visualized the learned global filters in the frequency domain to interpret the model's feature extraction process. By examining these visualizations, we aimed to identify distinct patterns in the frequency filters that could indicate which spatial or frequency components were prioritized during training. Clear, structured patterns in the global filters might suggest that the model has learned to emphasize certain frequency ranges, potentially associated with edges, textures, or other relevant image features. Conversely, a lack of clear patterns could imply that the model's filters were under-optimized or struggled to differentiate key features effectively.

Finally, training time and computational efficiency were evaluated alongside accuracy and loss, particularly to assess potential scalability advantages. The comparison of training time with self-attention models provided insights into the efficiency of the model. Together, these evaluations aimed to provide a comprehensive view of the potential of the model, guiding future improvements in both model architecture and training strategies.

Experimental Results

We present the findings of the model after testing, and some visualization to aid in the understanding of the model.

Results

The outputs for image accuracy and training time are shown in Table 1 for the model with and without the global filter on the CIFAR datasets. The global filter layer provided a marked increase in accuracy, with CIFAR10 accuracy improving by 23% and CIFAR100 accuracy improving by 14%. Additionally, the global filter layer reduced training time by about 25-27%, demonstrating its computational efficiency.

While the overall accuracies for the tested models are lower than those reported in the original paper, particularly on CIFAR100, this discrepancy may result from differences in hyperparameters, training epochs, and preprocessing methods. As mentioned previously, the original paper utilized 30 training epochs compared to the 10 epochs used in the current implementation, which along with other factors may have contributed to the lower accuracy. Further optimization or extended training times could potentially close this gap, especially given the complexity of CIFAR100.

Compared to the self-attention model used as a baseline, the GFNet model with a global filter showed both accuracy gains and faster training times on the CIFAR datasets, indicating the layer's potential as a computationally efficient alternative for image classification. These findings, while limited, suggest that GFNet with a global filter could be a viable option for applications requiring both speed and accuracy, and future research could explore its effectiveness on larger datasets and real-world tasks.

Model	Dataset	Accuracy	Time
Self Attention	CIFAR10	43.73%	779 s
Global Filter	CIFAR10	66.6%	566 s
Self Attention	CIFAR100	23.27%	782 s
Global Filter	CIFAR100	37.41%	584 s

Table 1: Accuracy on images and time of training for the tested models using either a self-attention layer or a global filter layer.

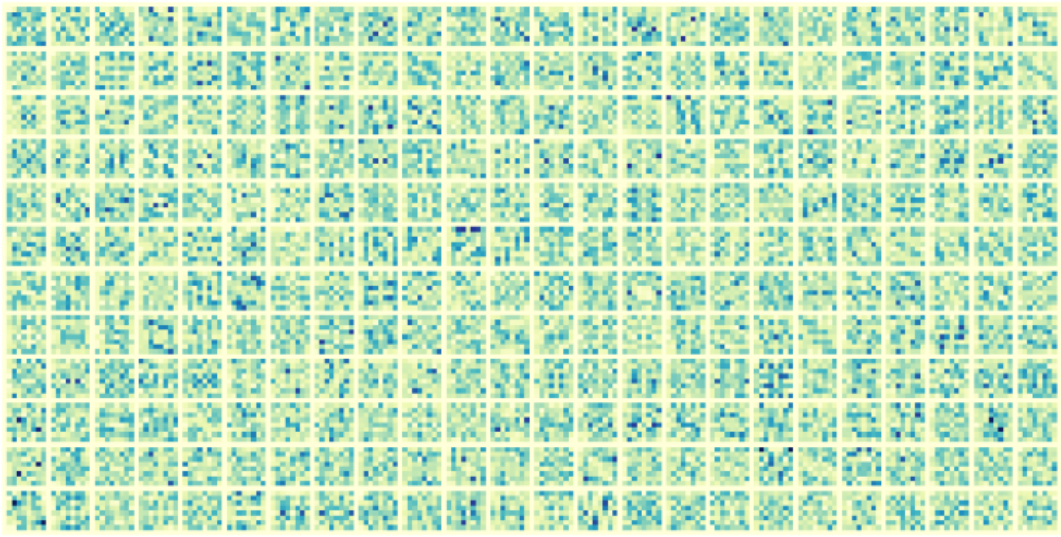

Figure 1: Visualization of the global filters in the implemented model. This shows the filters in the frequency domain, with each layer occupying a row, arranged with the top row representing the first layer, and the bottom row representing the twelfth and final layer.

Visualization

The main feature of the GFNet is the global filter in the frequency domain, which can be seen in Figure 1. Each row represents the filter layer that gets multiplied in the frequency domain after the Fourier transform. In interpreting the figure, we attempt to find distinct patterns across the filter layers, however, these patterns are less defined than those typically observed in highly optimized global filters. In frequency-based filtering, clear and structured patterns often indicate that the model has effectively learned to focus on specific frequency ranges, which correspond to spatial features like edges, textures, or broader structural elements within an image. The lack of these well-defined patterns in the current filters indicates that the model has not fully learned to differentiate meaningful frequency components effectively, which may be a reason for the low accuracies observed in the results.

Conclusion

In this project, through re-implementation, we explored the efficacy of GFNet models on the CIFAR10 and CIFAR100 datasets. While the model did not achieve the same levels of accuracy as reported in the original paper, the incorporation of the global filter layer demonstrated significant benefits in both accuracy and training efficiency. For CIFAR10 and CIFAR100, GFNet with the global filter achieved marked accuracy improvements, as detailed in the Results section.

The results indicate that GFNet's global filter could be a valuable addition to image classification architectures, particularly as a computationally efficient alternative. The ability of the global filter to boost performance without requiring deeper layers highlights its potential for use in resource-constrained applications.

Future Directions

The work presented in this paper was more simple and conceptual, so there remains much more potential in researching this topic. Future work can build on these findings in a large variety of ways.

Looking specifically at the model in the implementation, there are many opportunities to contrive more useful results beyond those in this paper.

- Additional Metrics: While only accuracy and time of training were considered in this paper, there are many other important parts that determine the value of a model. Some metrics that could be further examined include floating point operations per second (FLOPS), parameters, latency, and GPU memory, which are each important in evaluating the complexity of the model. Robustness and generalization ability could also be looked into as well, given their importance in practical efficacy.

- Optimization and Tuning: Fine-tuning hyperparameters, exploring alternative learning rates, and training for longer epochs could help bridge the accuracy gap between our implementation and the results reported in the original paper. The original paper itself had a multitude of variants for their model. Additional training on CIFAR100, in particular, could help address the dataset's complexity and further improve accuracy.
- Comparison to More Models: While the self attention model being compared to in this paper were chosen for direct comparison in order to determine exactly the impact of the global filter layer, testing and comparison to other types of models would allow for a clearer understanding of the place of GFNet among image classifiers as a whole. There could even be further testing done still on the layer level to find if there are other alternatives that outperform the global filter as well.
- Expanding to Larger Datasets: Applying GFNet with the global filter layer to larger and more diverse datasets, such as ImageNet, would provide insights into the scalability of the model. This could validate the global filter's ability to generalize across more complex image classification tasks. This ability is one of the main points of GFNet given its log linear complexity. Applying GFNet to multiple datasets would be important in analyzing its performance compared to increasing token size and complexity.

Beyond the model in this paper, there is still much to explore with global filters, with the following detailing some avenues of further development in new models and other applications.

- Different Architectures: Experimenting with hybrid architectures that combine the global filter layer with self-attention or transformer-based modules could enhance the model's ability to capture both global and local dependencies. This hybrid approach may yield better performance, especially on tasks requiring high detail resolution. In addition, the hierarchical version of the model could also be looked in to as a possibility to improve performance. Further research could determine where the global filter would best apply.
- Real-World Applications: Testing GFNet in real-world applications could highlight the practical benefits of the global filter layer in scenarios requiring efficient computation, and is most important in practical use.
- Investigating Frequency Domain Variants: Exploring alternative methods of frequency domain filtering, such as dynamic or adaptive filters that change based on input characteristics, could further improve GFNet's performance. These approaches might offer more flexible control over feature extraction, optimizing accuracy across various types of image data.

In summary, the project demonstrates the promise of GFNet's global filter layer in improving classification accuracy and efficiency. Future research can expand these findings to validate GFNet's robustness and explore novel architectural variations, potentially positioning GFNet as an efficient choice for modern image classification tasks.

References

Chi, L.; Jiang, B.; and Mu, Y. 2020. Fast Fourier Convolution. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 4479–4488. Curran Associates, Inc.

Rao, Y.; Zhao, W.; Zhu, Z.; Lu, J.; and Zhou, J. 2021. Global Filter Networks for Image Classification. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 980–993. Curran Associates, Inc.