WAVEMIX: MULTI-RESOLUTION TOKEN MIXING FOR IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Even though vision transformers (ViTs) have provided state-of-the-art results on image classification, their requirements of large data, model size, and GPU usage have put them out of reach of most practitioners of computer vision. We present WaveMix as an alternative to self-attention mechanisms in ViT and convolutional neural networks to significantly reduce computational costs and memory footprint without compromising on image classification accuracy. WaveMix uses a multi-level two-dimensional discrete wavelet transform for mixing tokens and aggregating multi-resolution pixel information over long distances, which gives it the following advantages. Firstly, unlike the self-attention mechanism of ViT, WaveMix does not unroll the image. Thus, it has the right inductive bias to utilize the 2-D structure of an image, which reduces the demand for large training data. Additionally, the quadratic complexity with respect to sequence length is also eliminated. Secondly, due to its multi-resolution token-mixing, WaveMix also requires much fewer layers than a CNN does for comparable accuracy. Preliminary results from our experiments on supervised learning using CIFAR-10 dataset show that a four-layer WaveMix model can be 37% more accurate than a ViT with a comparable number of parameters, while consuming only 3% of the latter's GPU RAM and memory. This model also performs better than efficient transformers and models not based on attention, such as, FNet, and MLP Mixer. Scaling up the WaveMix model to achieve a top-1 accuracy of over 85% on CIFAR-10 could be done on a 16 GB GPU, while consuming only 6% of the GPU RAM used by the largest ViT which could fit in that GPU. Our work suggests that research on model structures that exploit the right inductive bias is far from over, and that such models can enable the training of computer vision models in settings with limited GPU resources.

1 INTRODUCTION

The self-attention mechanism in the transformer architecture (Vaswani et al., 2017) has been successful in utilising long range relationships between tokens and achieving state-of-the-art results in NLP and computer vision tasks (Dosovitskiy et al., 2021). However, in low data regime, transformer models are shown to perform poorly compared to convolutional models as they lack the proper inductive bias, and hence require more data to model the 2D image features. This demand for huge data, especially for pre-training, limits the use of vision transformer (ViT) models among practitioners who work with low data. Also, the quadratic complexity of self-attention with respect to sequence length, which can be large for image data (number of pixels), creates another challenge for the training of ViT models. These limitations of self-attention have sparked research into finding better alternatives that can find the long range relationships between tokens with reduced data and computational costs. Most of the work towards reducing the complexity of self-attention has been in creating sparse attention models and linear attention mechanisms that can approximate the attention matrix. Alternatives, such as Hybrid Vision X-formers (Jeevan & Sethi, 2021), utilise the inductive priors, such as convolutions, and linear attention mechanisms, such as Performer and Nyströmformer, to reduce computational costs (Choromanski et al., 2021; Xiong et al., 2021).

We propose WaveMix as an alternative neural network architecture for image analysis that can achieve performance comparable to self-attention models while consuming orders of magnitude fewer GPU memory, model size, and computations for the same accuracy as a ViT architecture.

WaveMix uses a two-dimensional discrete wavelet transform for token mixing. WaveMix is more suitable for handling images than ViT architectures because it does not unroll the image at any point. We demonstrate that WaveMix can scale up and work on low resource environments (GPU $\leq$ 16GB) with less data, while still giving higher accuracy in image classification. Compared to the standard ViT, WaveMix performed 37% better on CIFAR-10 dataset, while using only 3% of the GPU RAM as that used by ViT. WaveMix combines the long range token mixing of self-attention, and efficiency, both computationally and in terms of memory, of CNNs.

1.1 Related Works

Experiments have shown that replacing the self-attention in transformers with fixed token mixing mechanisms, such as the Fourier transform, achieves comparable performance at lower memory and computational costs (Lee-Thorp et al., 2021). This has led to search for other linear transforms that can mix the tokens in a way that helps the model learn the inter-dependencies between them. Given the multi-resolution analysis properties of the wavelet transform that are suitable for natural images, which have been exploited for image denoising (Ruikar & Doye, 2010), super-resolution (Guo et al., 2017), recognition (Mahmood et al., 2018), and compression (Lewis & Knowles, 1992), we propose using the two-dimensional Discrete Wavelet Transform (2D DWT) for token mixing.

Among the different types of mother wavelets available, we used the Haar wavelet, (a special case of the Daubechies wavelet (Daubechies, 1990)) also known as Db1 which is frequently used due to its simplicity and faster computation. Haar wavelet is both orthogonal and symmetric in nature, and have been used to extract basic structural information from images (Porwik & Lisowska, 2004).

2D DWT for extracting image features has been used extensively in machine learning literature (Ghazali et al., 2007). Most of the previous work focused on using 2D DWT in conjunction with other machine learning models, such as support vector machines and neural networks, especially for classification of medical images (Ranaware & Deshpande, 2016; Nayak et al., 2016). ScatNet architecture cascades wavelet transform layers with non-linear modulus and average pooling to extract a translation invariant feature that is robust to deformations and preserves high-frequency information for image classification (Bruna & Mallat, 2013). WaveCNets replaces max-pooling, strided-convolution, and average-pooling of CNNs with DWT for noise-robust image classification (Li et al., 2020a). Multi-level wavelet CNN (MWCNN) is used for image restoration in U-Net architectures for better trade-off between receptive field size and computational efficiency (Liu et al., 2018). Wavelet transform has also been combined with a fully convolutional neural network for image super resolution (Kumar et al., 2017).

Our work extends a large body of literature that uses wavelet transforms in neural networks. Rather than using the wavelet transform to extract image features for use in a downstream neural network, we build wavelet blocks that process images at multiple resolutions, and then combine the information from this multi-resolution analysis to gather relationships between tokens. Using a linear transform, such as the DWT, reduces the model size considerably, since there are no learnable parameters in the DWT operation.

2 WaveMix Architecture

Our model consists of a series of WaveMix blocks that perform 2D DWT on the input and extract the approximation and detail coefficients. The input image is first passed through a convolutional layer that creates feature maps of the image, as shown in Figure 1. Convolutional layers have the ability to learn representations of low level image features in the earlier layers more efficiently due to their strong inductive bias (Graham et al., 2021). The number of feature maps generated is equal to the dimension of embedding needed. The feature maps generated by the CNN layers are passed to the WaveMix block where multi-level 2D DWT is applied and passed through feed-forward layers before it is send to the succeeding WaveMix blocks. A residual connection is provided within each WaveMix block so that the model can be made deeper with a larger number of blocks if necessary (He et al., 2015). The output from the last WaveMix block is then passed through a pooling layer and finally to an MLP head, which gives the output.

At no point in the token-mixing phase of the model do we unroll the image into a sequence of pixels. So we have developed a model that can exchange information between pixels which are separated by

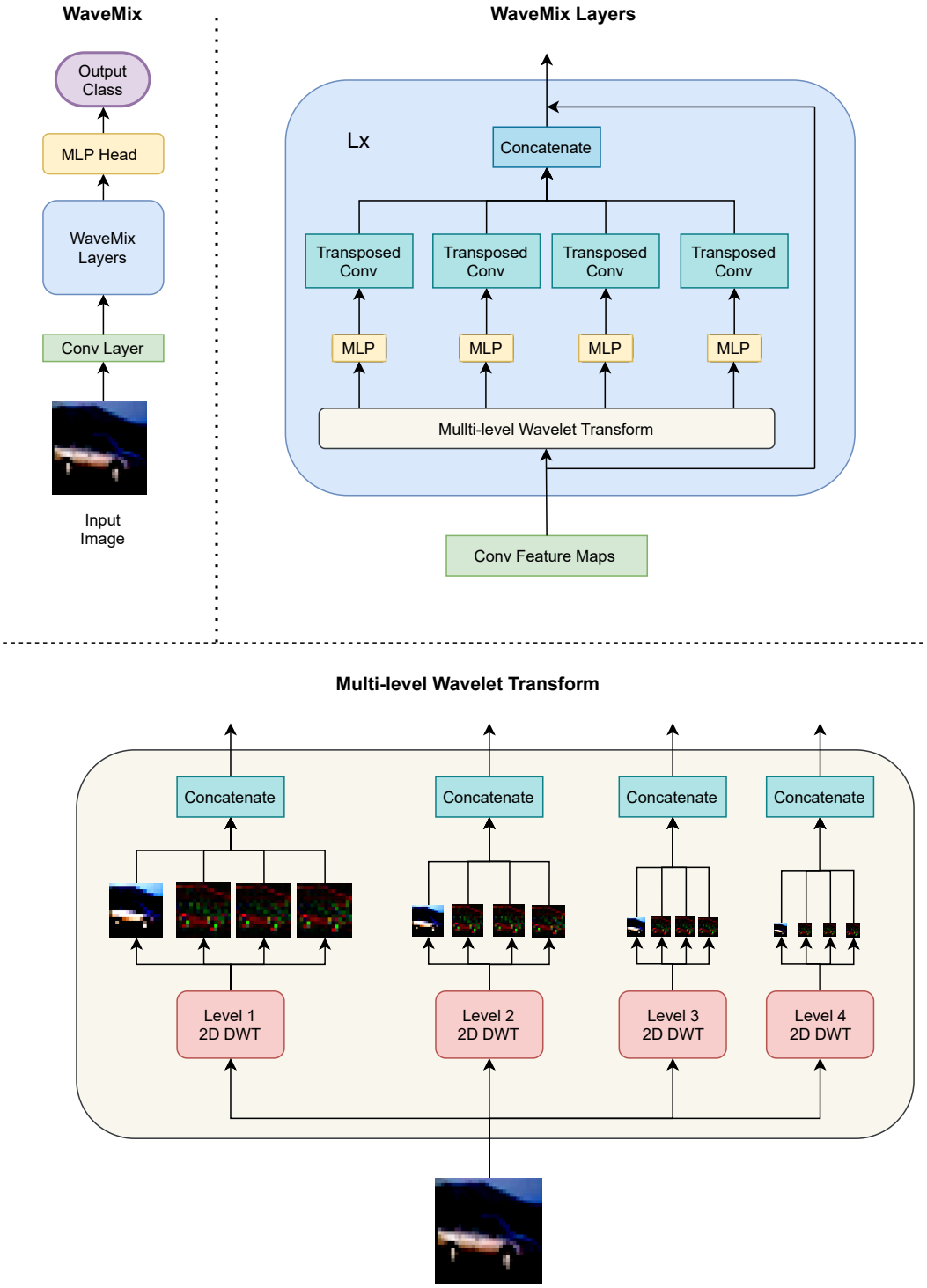

Figure 1: WaveMix Architecture

long distances without using self-attention and thereby escaping the quadratic complexity bottleneck of self-attention.

2.1 WAVEMIX BLOCK

The detail and approximation coefficients are extracted from the input using multi-level 2D DWT. We use Haar wavelet (Db1) for generating the 2D DWT output [1]. The number of levels needed is decided based on the image size. Each level reduces the DWT output by half, therefore, we use as many levels as necessary till the input size is reduced to 2×2. For example, a 32×32 image requires a 4-level 2D DWT, which creates 16×16, 8×8, 4×4 and 2×2 sized outputs respectively. We retain the low resolution images generated at each level and concatenate the corresponding 3 sets of approximation coefficient outputs from the same level of 2D DWT along the channel dimension, as shown in Figure 1.

The concatenated output from each DWT level is passed through feed-forward layers to reduce the feature maps and passed through deconvolutional layers. The kernel size and strides are chosen such that all the different sized outputs from different levels of DWT are brought back to the image size. These output feature maps are then passed through another convolutional layer which reduces the number of feature maps such that when output of this layer is concatenated, it creates as many feature maps as was input to the WaveMix block.

Each level of 2D DWT creates a low-resolution approximation of the image with the information on reconstructing the original high-resolution image in separate channels. The first and second level resolutions capture the finer details of the image while further lower-stage resolutions capture more global information. The feed-forward sub-layers immediately following this DWT have access only to the outputs at the corresponding level to learn the features. Once the features learned from each resolution level are passed through transposed convolutions, where all the different low-resolution images are up-sampled to full image size and concatenated along channel dimension, the feed-forward sub-layers in the succeeding layers have full access to all the local and global information carried by the tokens. Transposed convolution of the lower resolution DWT outputs will spread the global information to all regions of the image which helps the succeeding feed-forward sub-layers understand relationships between tokens both locally and globally.

3 EXPERIMENTS

The CIFAR-10 dataset was used for the experiments (Krizhevsky, 2009). Models were also trained on the Tiny ImageNet dataset to analyse their performance on larger images (Le & Yang, 2015). All the 2D DWT computations were done using Haar wavelet. The Adam optimizer with $\alpha = 0.001$ (learning rate), $\beta_1 = 0.9$, and $\beta_2 = 0.999$ was used for computing running averages of gradient and its square, with $\epsilon = 10^{-8}$ and a weight decay coefficient of 0.01. We used automatic mixed precision in PyTorch during training to make it faster and consume less memory. Experiments were done with 16 GB Tesla P100-PCIe and Tesla T4 GPU available in Kaggle and Google Colab. GPU usage for a batch size of 64 was reported along with top-1% and top-5% accuracy from best of 3 runs. Patch size of 1 was chosen for all the models that unrolled the images as a sequence of pixels, such as the ViT.

We applied 2 layers of 3×3 convolutions to the input image with stride and padding set to 1. The 2 CNN layers increased the channel dimension from 3 to the required final embedding dimension in 2 stages. We used linear and convolutional feed-forward layers with 1×1 kernel in our experiments, as the number of parameters of linear layer and 1×1 convolutional layer was almost the same, and using larger kernel sizes would have significantly increased the number of parameters in the network. Performance was tested by increasing the number of feature maps (embedding dimensions) and varying the depth and dropout rates.

[1] The code for 2D DWT was taken from Pytorch Wavelets

Table 1: Performance of different models on CIFAR-10 dataset

Models	Parameters	Size (MB)	Top-1 Accuracy	Top-5 Accuracy	GPU (GB)
ViT	0.53 M	209.39	57.39	94.98	14.1
Hybrid ViN	0.61 M	88.38	75.26	98.39	5.0
FNet	**0.27 M**	60.12	51.54	93.84	1.6
MLP Mixer	8.53 M	90.59	60.33	95.79	1.4
WaveMix	0.57 M	**6.24**	**78.55**	**98.72**	**0.4**

Table 2: Performance of different WaveMix models on CIFAR-10 dataset

Models	Parameters	Size (MB)	Top-1 Accuracy	GPU (GB)
WaveMix-16	0.16 M	3.09	75.74	0.2
WaveMix-32	0.57 M	6.24	78.55	0.4
WaveMix-64	2.3 M	16.86	83.63	0.7
WaveMix-128	9.2 M	51.22	85.21	0.8

4 RESULTS

Table 1 shows the performance of 4-layer WaveMix compared to other 4-layer architectures for CIFAR-10 classification using supervised learning. We choose a WaveMix model having almost the same number of parameters as ViT for a better comparison of performance. We can see that WaveMix outperforms all other models, especially the ViT by 37% and Hybrid Vision Nyströmformer (ViN) by 4%. This shows that in the low data regime, WaveMix is a good alternative to attention-based architectures. It performs 53% better than FNet since we use the wavelet instead of the Fourier transform, where the former is better suited for multi-resolution modeling of the image data and does not need unrolling the images as a sequence of pixels.

WaveMix performed 236% better than a 4-layer CNN which was trained on this dataset. The higher accuracy obtained is due the ability of the WaveMix model to process the image at multiple resolutions in parallel where it can learn image features obtained at different scales. This ability is absent in convolutional layers, which require pooling for large-scale information mixing, and it comes at the cost of quadratic complexity in attention networks.

In our notation, we use the embedding dimension to differentiate various WaveMix models. The number following *WaveMix* is its embedding dimension; for example, WaveMix-64 has an embedding dimension of 64. The FNet has less parameters than WaveMix-32 as shown in Table 1 and is more comparable to WaveMix-16 shown in Table 2 with respect to the number of parameters. We see that even the smaller WaveMix-16 outperforms FNet by 47% using only 5% of its memory and 12% of its GPU RAM usage. FNet just uses one feed-forward sub-layer in each layer, while WaveMix uses multiple parallel feed-forward sub-layers and transposed convolution sub-layers in each layer. This helps WaveMix in mixing tokens at multiple resolutions to learn image representations better than the FNet.

Table 2 shows the performance and resource consumption of 4-layer WaveMix models differing in their embedding dimensions. We can see that increasing the embedding dimension increases the performance of the model without much increase in the GPU RAM consumption. These results demonstrate the ability of the WaveMix architecture to be extremely efficient in low-resource environments even while using high embedding dimensions.

The WaveMix architecture uses only $\frac{1}{35}$-th of the GPU RAM consumed by a ViT that has a similar number of parameters. More importantly, the WaveMix and can still outperform the ViT in terms of accuracy when used in classification of small datasets, such as CIFAR-10 and Tiny ImageNet, as shown in Figure 2. Since the major limitation with use of self-attention based models is need for large GPU resources and memory, WaveMix model offers a new way forward in the search for low-GPU models which can still provide global token-mixing similar to attention.

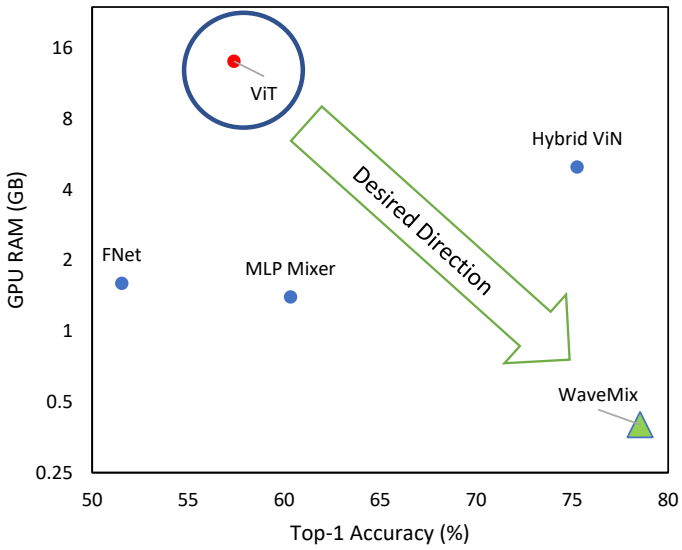

Figure 2: Comparison of classification accuracy and GPU usage by various models on CIFAR-10 dataset for a batch size of 64

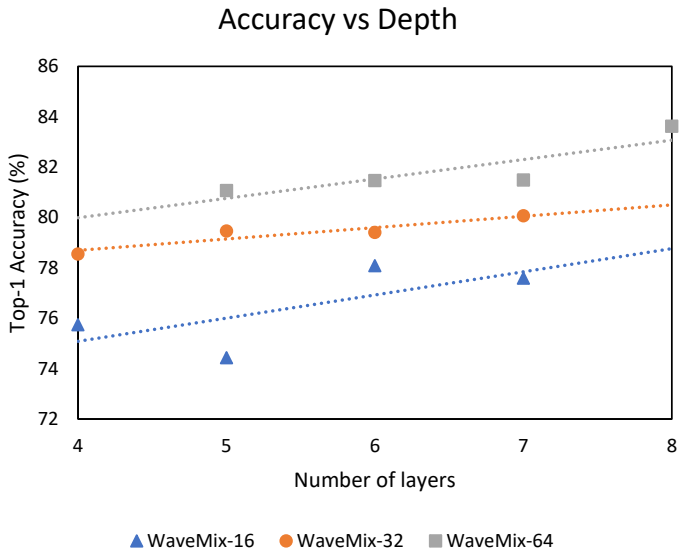

Figure 3: Variation of accuracy with depth for various WaveMix models

There is also a higher flexibility with the WaveMix architecture in controlling the number of parameters, as we can separately change the embedding dimensions of the feature maps and output of the feed-forward sub-layers, since the embedding dimension of output of the WaveMix layer is only dependent on how the concatenation of transposed convolutions are performed.

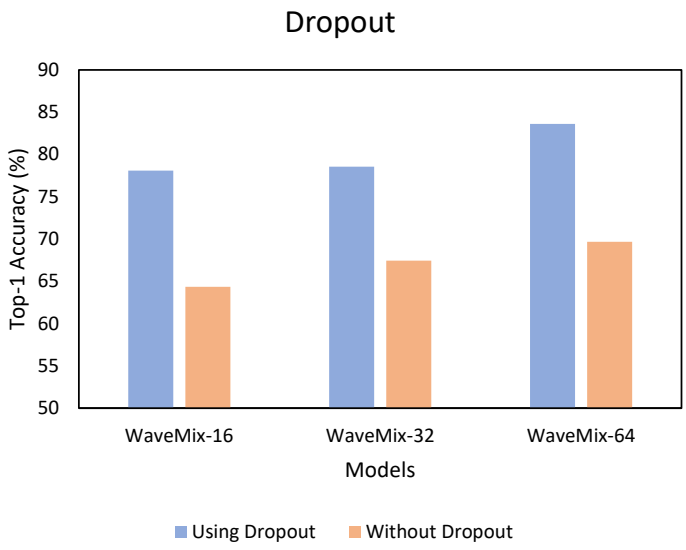

Figure 4: Impact of Dropout in various WaveMix Models

Figure 3 shows the variation of accuracy in image classification with depth for different WaveMix models. We observe the general trend where accuracy of the model increases with addition of each WaveMix layer across different model embedding sizes.

Our experiments also show that dropout rates have a significant impact on the accuracy of the models as shown in Figure 4. Low dropout rates (0.2-0.3) were found to be optimal for smaller models while higher dropout rates (0.5-0.6) were found to be better for larger models. This huge improvement in accuracy (greater than 10 percentage points) is surprising since dropout does not cause such a large performance boost in either transformers or CNN networks.

We also checked the importance of convolutional layers by dropping the initial two convolutional layers before the WaveMix blocks. The performance deteriorated by more than 10% without these layers. This shows that the convolutional layers provides the inductive bias to learn low-level image features, such as edges, which are useful for the global token mixing performed by the WaveMix blocks.

We tried both, convolutional layer with 1×1 kernel, and linear layers in our feed-forward sub-layers and found that convolutional layers are better and faster. The main purpose of the feed-forward sub-layers is to learn the image features and reduce the embedding dimensions after concatenation.

4.1 Higher Resolution Images

For the classification of higher resolution images of size 64×64 in the Tiny ImageNet dataset, we tried 3 different methods:

1. WaveMix ConvStride-2: Use the initial Convolution layer with stride 2, which will reduce the image output to size 32×32 and the WaveMix blocks will remain unchanged.
2. WaveMix Top-4 DWT: Use the first 4 levels of 2D DWT and remove the 5th level. This will create output sizes of 32×32, 16×16, 8×8, and 4×4 while we eliminate the 2×2 output.
3. WaveMix All-5 DWT: Use all 5 levels of 2D DWT but reduce the channel dimension of the 5th level so that the output channel dimension after concatenation matches the input in the WaveMix block.

Table 3: Performance of different models on Tiny ImageNet dataset

Models	Parameters	Size (MB)	Top-1 Accuracy	GPU (GB)
ViT	0.55 M	209.59	26.43	14.1
WaveMix ConvStride-2	0.58 M	**6.30**	31.57	**0.4**
WaveMix Top-4 DWT	0.58 M	**6.30**	**31.90**	**0.4**
WaveMix All-5 DWT	1.1 M	8.41	30.54	**0.4**

The comparison of performance of the three approaches are shown in Table 3, where we used a 4-layered WaveMix-32 to train on Tiny ImageNet. We see that removing the 2×2 resolution level does not affect the performance, and also does not increase the model size. All 3 models consumed the same amount of GPU RAM but the model which used all 5 levels of 2D DWT had almost twice the number of parameters and used slightly more memory than the others. For tasks involving higher resolution images of sizes 256×256 or 512×512, we recommend using at least 6 level 2D DWT. We also see that WaveMix models outperform ViT by 17% for classification of the Tiny ImageNet dataset.

5 Conclusions

The WaveMix architectures offers the best of both self-attention networks and CNNs by combining long distance token mixing of attention, and low GPU consumption, efficiency and speed of CNNs. It is more tailored to computer vision applications as it handles the data in 2D format without unrolling it as sequence as done by the ViT. Experiments on low data image classification tasks show that WaveMix achieves considerably higher accuracy with less than 3% of the GPU RAM when compared to a ViT with comparable number of parameters.

More testing is needed to analyse the performance of WaveMix in classification of higher resolution images and also in other computer vision tasks, such as object detection and segmentation. Fusing this architectures with efficient attention layers might also be a way forward. We should also test different mother wavelets for various computer vision tasks and find which family of wavelets is suitable for each task. Identifying the right wavelet may significantly improve the performance of WaveMix by introducing an even more optimal inductive bias and sparseness of response. It is possible that the output of the convolutional filters require a different wavelet family compared to those required for the pixel for optimal performance. Additionally, the role of sparseness of wavelet response needs to be studied as well, which was found to be desirable for image compression, denoising, and pattern recognition. Further experiments with hyper-parameter tuning, alternative mixing models, image datasets and vision tasks can lead to more insights for improving the accuracy and efficiency in low data, low GPU RAM regime.

Our research suggests several significant directions for developing alternatives to CNN and attention based architectures for vision. While transformer architectures have produced some of the highest image classification accuracies, they come with higher costs in terms of training data, computations, GPU RAM, hardware costs, and power consumption Li et al. (2020b). This makes training attention-based architectures inaccessible for most practitioners, except those who work for a select few organizations with massive resources. On the other hand, we also see that proposals of neural architectures that exploit domain-specific inductive biases have resulted in usable increases in performance and decreases in computational and data requirements. Our work suggests that the quest for better, efficient, faster and smaller models is far from over, and these innovations can democratize the ability to train neural networks from scratch to state-of-the-art performance.

Acknowledgments

We would like to thank Dr. Neeraj Kumar from University of Alberta for the his useful insights on the wavelet Transform and its application to images.

REFERENCES

Joan Bruna and Stephane Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, 2013. doi: 10.1109/TPAMI.2012.230.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers, 2021.

I. Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36(5):961–1005, 1990. doi: 10.1109/18.57199.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Kamarul Hawari Ghazali, Mohd Fais Mansor, Mohd. Marzuki Mustafa, and Aini Hussain. Feature extraction technique using discrete wavelet transform for image classification. In *2007 5th Student Conference on Research and Development*, pp. 1–4, 2007. doi: 10.1109/SCORED.2007.4451366.

Ben Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference, 2021.

Tiantong Guo, Hojjat Seyed Mousavi, Tiep Huu Vu, and Vishal Monga. Deep wavelet prediction for image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1100–1109, 2017. doi: 10.1109/CVPRW.2017.148.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Pranav Jeevan and Amit Sethi. Vision xformers: Efficient attention for image classification, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Neeraj Kumar, Ruchika Verma, and Amit Sethi. Convolutional neural networks for wavelet domain super resolution. *Pattern Recognition Letters*, 90:65–71, 2017.

Ya Le and X. Yang. Tiny imagenet visual recognition challenge. 2015.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms, 2021.

A.S. Lewis and G. Knowles. Image compression using the 2-d wavelet transform. *IEEE Transactions on Image Processing*, 1(2):244–250, 1992. doi: 10.1109/83.136601.

Qiufu Li, Linlin Shen, Sheng Guo, and Zhihui Lai. Wavelet integrated cnns for noise-robust image classification, 2020a.

Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, and Joey Gonzalez. Train big, then compress: Rethinking model size for efficient training and inference of transformers. In *International Conference on Machine Learning*, pp. 5958–5968. PMLR, 2020b.

Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration, 2018.

Maria Mahmood, Ahmad Jalal, and Hawke A. Evans. Facial expression recognition in image sequences using 1d transform and gabor wavelet transform. In *2018 International Conference on Applied and Engineering Mathematics (ICAEM)*, pp. 1–6, 2018. doi: 10.1109/ICAEM.2018.8536280.

Deepak Ranjan Nayak, Ratnakar Dash, and Banshidhar Majhi. Brain mr image classification using two-dimensional discrete wavelet transform and adaboost with random forests. *Neurocomputing*, 177:188–197, 2016.

Piotr Porwik and Agnieszka Lisowska. The haar-wavelet transform in digital image processing: Its status and achievements. *Machine graphics & vision*, 13:79–98, 2004.

Preeti N. Ranaware and Rohini A. Deshpande. Detection of arrhythmia based on discrete wavelet transform using artificial neural network and support vector machine. In *2016 International Conference on Communication and Signal Processing (ICCSP)*, pp. 1767–1770, 2016. doi: 10.1109/ICCSP.2016.7754470.

Sachin Ruikar and D D Doye. Image denoising using wavelet transform. In *2010 International Conference on Mechanical and Electrical Technology*, pp. 509–515, 2010. doi: 10.1109/ICMET.2010.5598411.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention, 2021.