Dynamic Circular Convolution for Image Classification

Xuan-Thuy Vo, Duy-Linh Nguyen, Adri Priadana, and Kang-Hyun Jo

Department of Electrical, Electronic and Computer Engineering,
University of Ulsan, Ulsan (44610), South Korea
Email: xthuy@islab.ulsan.ac.kr;
{ndlinh301,priadana3202}@mail.ulsan.ac.kr; acejo@ulsan.ac.kr

Abstract. In recent years, Vision Transformer (ViT) has achieved an outstanding landmark in disentangling diverse information of visual inputs, superseding traditional Convolutional Neural Networks (CNNs). Although CNNs have strong inductive biases such as translation equivariance and relative positions, they require deep layers to model long-range dependencies in input data. This strategy results in high model complexity. Compared to CNNs, ViT can extract global features even in earlier layers through token-to-token interactions without considering geometric location of pixels. Therefore, ViT models are data-efficient and data-hungry, in another work, learning data-dependent and producing high performances on large-scale datasets. Nonetheless, ViT has quadratic complexity with the length of the input token because of the natural dot product between query and key matrices. Different from ViTs-and-CNNs-based models, this paper proposes a Dynamic Circular Convolution Network (DCCNet) that learns token-to-token interactions in Fourier domain, relaxing model complexity to $O(Nlog(N)$ instead of $O(N^2)$ in ViTs, and global Fourier filters are treated dependently and dynamically rather than independent and static weights in conventional operators. The token features, dynamic filters in spatial domain are transformed to frequency domain via Fast Fourier Transform (FFT). Dynamic circular convolution, in lieu of matrix multiplication in Fourier domain, between Fourier features and transformed filters are performed in a separable way along channel dimension. The output of circular convolution is revered back to spatial domain by Inverse Fast Fourier Transform (IFFT). Extensive experiments are conducted and evalued on large-scaled dataset ImageNet1k and small dataset CIFAR100. On ImageNet1k, the proposed model achieves 75.4% top-1 accuracy and 92.6% top-5 accuracy with the budget 7.5M paramaters under similar setting with ViT-based models, surpassing ViT and its variants. When fine-tuning the model on smaller dataset, DCCNet still works well and gets the state-of-the-art performances. Both evaluating the model on large and small datasets verifies the effectiveness and generalization capabilities of the proposed method.

Keywords: Vision Transformer · Dynamic Global Weights · Fourier Transform · Image Classification

1 Introduction

In the view of understanding involved visual data, the model compresses high dimension of image data to lower spaces and keeps informative features through processing layer-by-layer of the model. The way the model compresses and extracts the features relies on what the image encompasses. As we interpret datas, one point in the image contains two components: content (intensity values) $c \in \mathbb{R}^3$ and geometric information $w \in \mathbb{R}^2$. The image is interpreted as $I \in \mathbb{R}^{5\times N}$, where $N = H \times W$ is number of pixels in the image. With the formulation of convolution, CNNs aggregate information of local windows to the center of the local windows in the sliding manner and also capture the relative position w_{i-j} inside local window. General speaking, CNN models [8, 13, 20] can extract helpful features that the image contains and result in translation equivariance and locality. Otherwise, Transformer invented by [32] views a sentence as a sequence of words (tokens) and compute word-to-word relationship and dynamically aggregate these features by global multi-head self-attention blocks for machine translation. With the success of Transformer in both general modeling capabilities and scalable models, ViT [5] tries to adapt self-attention operation in computer vision. Each image is separated into a sequence of patches (tokens), and the model learns an affinity matrix of token-to-token similarity. The ViT only considers content-to-content relationships from the input images or input features and can fail to capture positional information. The lack of geometric w_{i-j} results in weak inductive biases. The model needs a lot of data to compensate for the absence of w_{i-j}.

In terms of model complexity, the convolution operation is more efficient than the self-attention block. To extract global features, CNN-based models stack a series of convolution layers with residual connections that create a large computational cost. At the heart of Transformer, self-attention operation requires quadratic complexity with the lengths of input tokens and the model is not acceptable to adapt self-attention operation at earlier layers. Especially for down-stream tasks, these networks perform predictions on the input features with high resolution. With the bottleneck computation of ViT, many methods try to reduce the cost $O(N^2)$ to $O(N)$ [22], sub-sample the query, key, and value matrices [33, 34], and compute attention in local windows mimicking convolution [18, 19]. Another line of research is to enhance the weak inductive biases of the transformer. The affinity matrix is supplemented with positional information such as absolute positional embedding [32], relative positional embedding [2, 4, 19, 23]. Other works [14, 15, 21, 22] attempt to combine the strengths of convolution and self-attention operations to build hybrid networks. They inherit the strong inductive biases of CNNs and the strong modeling of ViTs, and deliver better performance than pure CNNs and ViTs.

On the research trend of Transformer, this paper develops a new operator, dubbed Dynamic Circular Convolution (DCC), which can extract and aggregate global features by performing the circular convolution between reweighted global Fourier kernels and Fourier transformed features. The reweighting coefficients are generated conditioned on the input features and are dynamically adopted

according to the content of the input. The DCC layers are used to replace self-attention blocks in ViT, called DCCNet. Our proposed DCCNet brings four benefits: (1) Global features are extracted in one layer; (2) the content and geometric information of the input image are utilized when computing circular convolution; (3) the generated weights are input-dependent instead of input-independent in conventional convolution; and (4) the complexity is $O(N(logN))$ rather than $O(N^2)$ in Transformer.

To verify the effectiveness of the proposed method, we conduct the experiments on the large dataset Imagenet1k, and small dataset CIFAR100. As a result, the DCCNet surpasses the baseline ViT and its variant by a clear margin under the same setting and budget (7.5M parameters and 1.2 GFLOPs).

2 Related Works

In this section, we briefly review some related works about Convolutional Neural Networks, Vision Transformer and its variant, and Fourier transform in computer vision.

CNNs: In 2012, with the development of parallel hardware computation, AlexNet [13] successfully train the convolution networks on large datasets and open new directions in Computer Vision. VGG [28] enlarge the network depth by stacking a sequence of plain 3×3 convolutions. Even though VGGNet achieves the large improvement on large-scale ImageNet dataset, the model causes vanishing gradient problem when the depth beyond 19 layers. ResNet [8] proposes residual blocks that can eliminate vanishing gradient and number of layers are stacked up to 1000 layers. From that event, many works are introduced to improve the baseline ResNet such as dense connection [11], deformable convolution [3], depthwise seperable convolution [10, 27], and multiple branches [30].

ViT: Recently, ViT [5] integrated the original Transformer [32] developed for natural language processing and established new state-of-the-art performances on image classification and downstream tasks. Because ViT has a simple structure and uniform representation, there are a lot of works that improve ViT model in both learning and cost. PVT [33] builds a multi-scale vision transformer network that gradually decreases spatial dimensions across stages. On each stage of PVT, key and value matrices are down-sampled to smaller token sizes. Instead of computing attention from all tokens, Swin [19] models local attention in predefined windows and also constructs hierarchical networks inspired by CNNs-based models. With these insightful properties, Swin outperforms the strong baseline ResNet [8] and sets new records in detection, segmentation and tracking performance. MobileViTv2 [22] proposes a separable self-attention operation that reduces the cost of original self-attention from $O(N)^2$ to $O(N)$.

Compensation for weak inductive biases of self-attention operation, methods [4,18,19,23] integrates relative positional information to attention maps. CPE [2] introduces a conditional positional encoding based on local relative neighborhood

of 3×3 depthwise convolution. Rather than marrying convolution operations to Transformer models, MobileViT [21] embeds Transformer blocks to stage 3, 4, 5 of MobileNetv2 [27]. Similar paradigm, NextViT [14] designs a hybrid network for embedded devices based on integration of the group convolution blocks in earlier stages and original self-attention blocks in later stages. EfficientFormer [15] adapt the idea of PoolFormer [38] and MetaFormer [39] and neural architecture search for constrained devices.

Based on the intuitive designs of Transformer and its variants, HorNet [24] extends matrix multiplication of self-attention operation to high-order interactions based on depth-wise separable convolution and recursive gates. FocalNet [37] uses multi-scale depth-wise separable convolutions and gated aggregation at each convolution to output multiple modulations.

Fourier Transform: FFC [1] proposes fast Fourier convolution and independently applies convolution and ReLU activation functions on the real and imaginary input features. Lama [29] adapt FFC operation to image inpainting. GFNet [25] learns global features in Fourier domain based on circular convolution and ViT models. AFNO [6] separates complex tensors into real and imaginary parts and utilizes the MLP module to mix these two parts. In this paper, we extend the circular convolution in GFNet to be dynamic and efficient. In GFNet, complex features and global filters are multiplied independently on each channel. Therefore, there is a way the model can efficiently learn the feature on both the spatial and channel axes. Moreover, in our core operator, both real and imaginary parts of the complex tensors are learned together instead of separation in the AFNO network.

3 Methodology

In this section, we leverage the overall network of the ViT [5] into our DCCNet in subsection 3.1 and analysis the proposed dynamic circular convolution block in subsection 3.2.

3.1 Overall Architecture

The DCCNet follows the single-scale architecture of the original ViT [5], shown in Fig. 1. Given input image with dimension $I \in \mathbb{R}^{3 \times H \times W}$, Patch Embedding splits and flattens the image I into a sequence of tokens with size $d_P \times N$, where N is the number of the tokens [1], H and W are height and width of image. Specifically, we use patch sizes of $P \times P$ and strides with patch window value over the image to produce the total tokens $N = \frac{H*W}{P^2}$ and $d_N = 3 * P^2$. Followed by non-overlap processing of the ViT implementation, 16×16 convolution with stride 16 is used in Patch Embedding as patch generation, corresponding to each token with size 16×16.

[1] Consistent term with original Transformer [32], also called number of patches

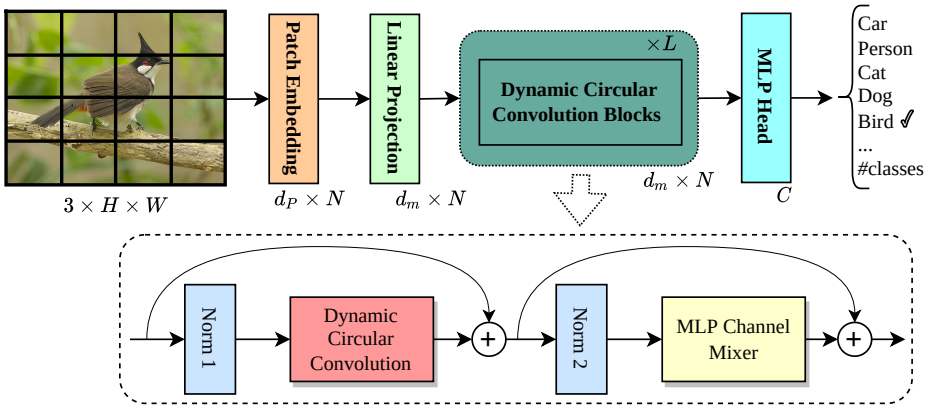


Fig. 1. The overall architecture of the DCCNet. N indicates the number of tokens with channel dimension d_m and L is the number of stacked DCC blocks. d_P, d_m are channel dimension after patch embedding, and channel dimension of the model. C is the number of predefined classes.

Linear Projection module projects a sequence of tokens with channel dimension d_P to a sequence of tokens with d_m. We use Linear layer to perform this process. The Dynamic Circular Convolution (DCC) block learns the token-token interaction that results in long-range dependencies between tokens. The DCC block includes two processes: (1) spatial mixings are performed by Dynamic Circular Convolution, and (2) MLP Channel Mixer mix token information along channel dimension. Between two processes, residual connections are used according to [38, 39] and each token is normalized by Layer Normalization before forwarding to each module. The detailed analysis of the DCC block is described in subsection 3.2. MLP Mixer contains two linear layers with expansion rate r. During training, based on [5, 25], we set $r = 4$ for all blocks.

Finally, Global Average Pooling (GAP) in MLP Head flattens a set of tokens to 1D dimension d_m and one linear layer projects flatten token d_m to number of classes C.

3.2 Dynamic Circular Convolution

The pipeline of the DCC operation is described in Fig. 2. Given the input feature $X \in \mathbb{R}^{d_m \times N}$, we reshape and permute the input X to 2D dimension $d_m \times H_P \times W_P$. Hence, the order of pixels in the input feature is still preserved. The permuted input features are processed in three steps: (1) 2D FFT (Fast Fourier Transform)transforms X in spatial domain to frequency domain by Fast Fourier Transform [1]; circular convolution is performed between transformed tensor and dynamic kernels to model global features; and 2D IFFT (Inverse Fast Fourier Transform) reserves dynamic and global tensor back to spatial domain.

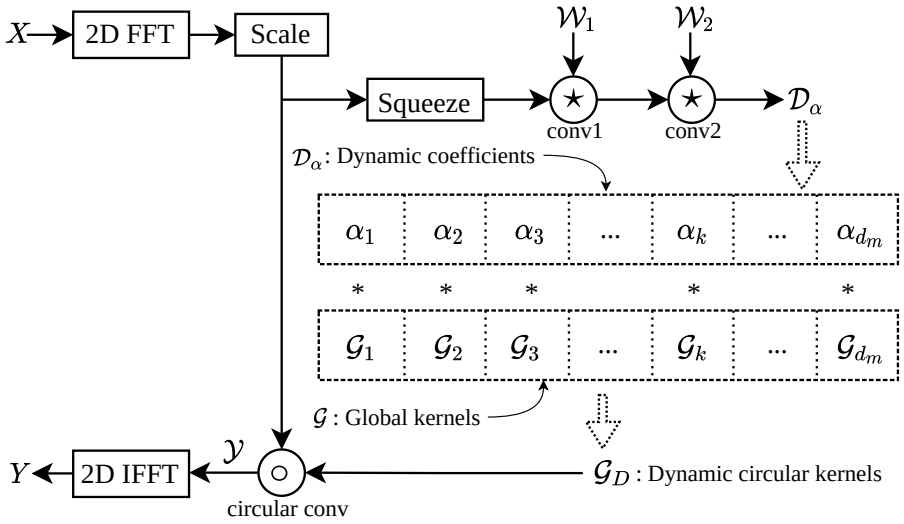


Fig. 2. The detailed architecture of the Dynamic Circular Convolution (DCC). 2D FFT and 2D IFFT denote Fast Fourier Transform and Inverse Fast Fourier Transform. $\mathcal{W}_1$, $\mathcal{W}_2$ are learnable parameters in frequency domain. Squeeze denotes mean computation along spatial dimension.

Given the permuted input with dimension $d_m \times H_P \times W_P$, complex tensor is generated by 2D FFT as follows,

$$\mathcal{X}[:,u,v] = \mathcal{F}(X) = \sum_{m}^{H_P-1} \sum_{n}^{W_P-1} X[:,m,n] e^{-j2\pi(\frac{um}{H_P}+\frac{vn}{W_P})}, \tag{1}$$

where $\mathcal{X}[:]$ is used to get the index of the channel. u, v are the coordinate of each output complex values $\mathcal{X} \in \mathbb{C}^{d_m \times H_P \times W_P}$ and m, n are the coordinate of each input real values $X \in \mathbb{R}^{d_m \times H_P \times W_P}$. $H_P = \frac{H}{P}, W_P = \frac{W}{P}$ are the height and width of the permuted sequence of tokens. Conventionally, angular frequencies along height and width dimensions are computed as:

$$\omega_h = 2\pi f_h = 2\pi \frac{u}{H_P}, \tag{2}$$

$$\omega_w = 2\pi f_w = 2\pi \frac{v}{W_P}. \tag{3}$$

In equation 1, there is a one-to-one mapping from the real domain to the frequency domain. It means that we convert the non-periodic signal to a periodic signal based on the theorem of the Fourier transform and fully preserve all the information of the input. The image can be decomposed into a function of *sine* waves.

One of the insightful property of Fourier transform is that there has a conjugate symmetry of the complex tensor $\mathcal{X}$ and leveraging such property can reduce the

model complexity without losing information [1]. This view can be represented as:

$$\mathcal{X}[:,u,v] = \mathcal{X}^*[:,H_P-u,W_P-v]. \tag{4}$$

Therefore, the model complexity is $O(H_P W_P log(H_P W_P))$ with respect to the length of the input tokens. While ViT-based models have quadratic complexity with the length of the input tokens, we enjoy much lower computational costs. A half of complex tensor $\mathcal{X}_s = \mathcal{X}[:,:,0:W_P/2+1]$ need to be computed and restored. It can relax memory intensive and still extract global features. Inside the equation 1, since the *sum* operation is used, the *scale* step is proposed to normalize all the accumulated values. During implementation, *scale* is conducted by average operation.

The model learns global features through self-attention operations or large kernel sizes. In this paper, we employ matrix multiplication between complex tensors $\mathcal{X}$ and global kernels. These global kernels are the same size as the scaled input $\mathcal{X}_s \in \mathbb{C}^{d_m \times H_P \times W_P/2+1}$. Hence, matrix multiplication between them in the spatial domain is called circular convolution in the frequency domain. In GFNet [25], they treat global kernels independently and statically because circular convolution is separable. It leads to a way that can mix the information of the input tensor along the channel dimension. Inspired by weight generation of self-attention operation [32], we define dynamic coefficients $\mathcal{D}_\alpha \in \mathbb{C}^{d_m \times 1 \times 1}$ as follows:

$$\mathcal{D}_\alpha = \{\alpha_1, ..., \alpha_{d_m}\} = \mathcal{W}_2 \star (\mathcal{W}_1 \star f(\mathcal{X}_s)), \tag{5}$$

where $\star$ is convolution operation. $f(.)$ indicates squeeze function that converts 2D input $\mathcal{X}_s$ to 1D vector. $\mathcal{W}_1 \in \mathbb{C}^{d_m \times \frac{d_m}{r}}$ and $\mathcal{W}_2 \in \mathbb{C}^{\frac{d_m}{r} \times d_m}$ are linear transformations in the frequency domain, mixing information of the squeezed complex tensor. Then, the dynamic coefficients $\mathcal{D}_\alpha$ is used to redistribute the static global kernel $\mathcal{G} \in \mathbb{C}^{d_m \times H \times (W/2+1)}$ via element-wise matrix multiplication,

$$\mathcal{G}_D = \{\alpha_i * \mathcal{G}_i | \alpha_i \in \mathbb{C}; \mathcal{G}_i \in \mathbb{C}^{H \times (W/2+1)}\} \in \mathbb{C}^{d_m \times H \times (W/2+1)}, \tag{6}$$

The dynamic circular kernel $\mathcal{G}_D$ is convoluted with the scaled input $\mathcal{X}_s$ to output global receptive field,

$$\mathcal{Y} = \mathcal{X}_s \circ \mathcal{G}_D, \tag{7}$$

where $\mathcal{Y} \in \mathbb{C}^{d_m \times H \times (W/2+1)}$ is the output of the circular convolution and $\circ$ denotes circular convolution.

Finally, we reserve the Fourier feature back to the spatial domain using the 2D Inverse Fast Fourier Transform (IFFT) and this operation is addressed as follows:

$$Y[:,m,n] = \mathcal{F}^{-1}(\mathcal{Y}) = \frac{1}{N} \sum_{u}^{H-1} \sum_{v}^{W-1} \mathcal{Y}[:,u,v] e^{j2\pi(\frac{um}{H} + \frac{vn}{W})}, \tag{8}$$

where N is the number of tokens used for normalization.

Table 1. Comparison with state-of-the-art models on ImageNet validation set

Method	Top-1 Acc (%)	Top-5 Acc (%)	#params	GFLOPs
T2T-ViT-7 [40]	71.7	-	4.3M	1.2
DeiT-Ti [31]	72.2	91.1	5.7M	1.3
gMLP-Ti [17]	72.3	-	7.0M	1.3
PiT-Ti [9]	73.0	-	4.9M	0.71
TNT-Ti [7]	73.9	91.9	6.1M	1.4
GFNet-Ti [25]	74.6	92.2	7.5M	1.3
LocalViT-T [16]	74.8	92.6	5.9M	1.3
ViTAE [36]	75.3	92.7	4.8M	1.5
DCCNet (our)	**75.5**	**92.7**	7.7M	1.2

4 Experiments and Results

4.1 Experiments

Datasets: The proposed DCCNet is trained and evaluated on the large-scale dataset ImageNet1k [26], and the small dataset CIFAR100 [12]. For ImageNet1k, this dataset includes 1.2M training images and 50k validation images with 1000 categories. CIFA10 contains 50k training and 10k testing images from 10 classes. Like CIFAR10, CIFAR100 contains 50k training and 10k testing images with 100 classes.

Experimental Setup: All implementations are conducted using the Pytorch framework, and the codebase is *Timm* [35] for fair comparisons with other methods. We follow the setting of methods [5, 25]. The model is trained for 300 epochs on two Tesla V100 GPUs. The batch size is 512 images per GPU, and the input images are resized to 224×224. The basic learning rate is $5 \times e^{-4}$ and learning schedule is cosine with warmup epochs of 5. The optimizer is AdamW with momentum 0.9 and weight decay 0.05. The DCCNet does not use EMA model and strong data augmentation like [19, 20].

4.2 Results

ImageNet dataset: Table 1 shows the main results evaluated on ImageNet validation set between DCCNet and other methods. As a result, DCCNet achieves 75.5% Top-1 accuracy and 92.2% Top-5 accuracy, which surpasses state-of-the-art ViT-based models around 7M parameters and 1.2 GFLOPs, such as 71.7% Top-1 in T2T [40], 72.2% in DeiT [31], 72.3% Top-1 in gMLP [17], 73.0% Top-1 in PiT [9], 73.9% Top-1 in TNT [7], 74.6% Top-1 in GFNet, 74.8% Top-1 in LocalViT [16], and 75.3% Top-1 in ViTAE [36]. This comparison verifies the effectiveness of the proposed DCCNet.

CIFAR 100: Table 2 describes the comparison between the DCCNet and other methods on the small dataset CIFAR100. With largely smaller parameters and GFLOPs than other methods, the DCCNet gets 84.1% Top-1 accuracy under budget 7.5M paramters and 1.2 GFLOPs.

Table 2. Results on small dataset CIFAR 100

Method	Top-1 Acc	#params	#GFLOPs
DeiT-T [31]	67.59	5.3M	0.4
PVT-T [33]	69.62	15.8	0.6
Swin-T [19]	78.07	27.5M	1.4
DCCNet (ours)	**84.10**	**7.5**	**1.2**

Ablation study: We investigate the effect of reduction ratio $r \in \{8, 16, 32\}$ in linear matrices of the DCC block on the model performance and cost illustrated in Table 3. When changing the reduction r, the Top-1 performances are similar. For a trade-off between accuracy and cost, we select $r = 16$ for all experiments.

Table 3. Ablation study on the reduction ratio r

Reduction r	Top-1 Acc (%)	Top-5 (Acc%)	#params	GFLOPs
8	75.4	92.7	7.9	1.3
16	**75.5**	**92.7**	7.7	1.2
32	75.2	92.4	7.6	1.2

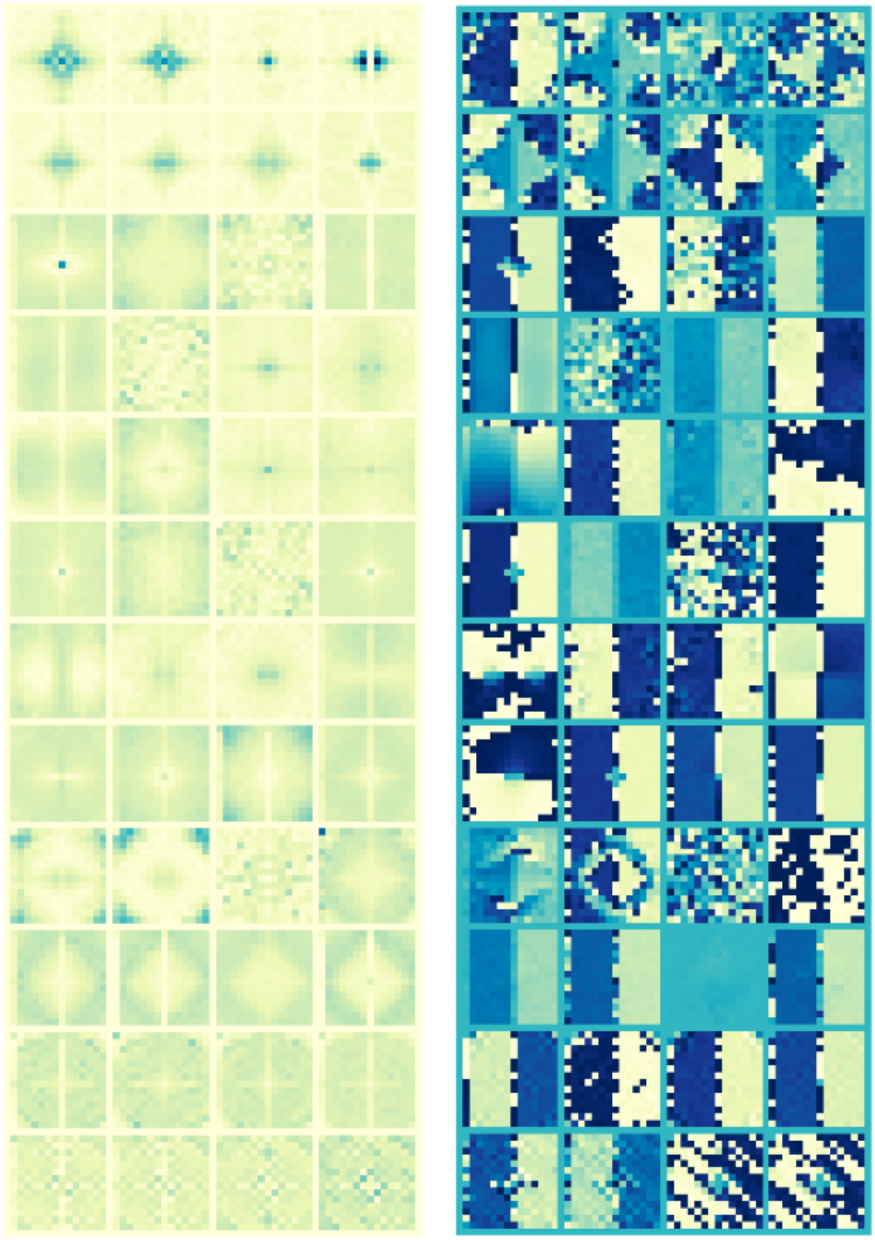
Amplitude and Phase Spectrum: We visualize the amplitude and phase spectrum on Fig. 3. As we can see, the detailed patterns in the amplitude spectrum are clear, and its spectrum has the symmetric property demonstrated in [1].

5 Conclusion

This paper presents a feature extractor based on Fast Fourier Transform and dynamic weight generations, called DCCNet. All spatial operations, especially for circular convolution, are performed in the frequency domain through the FFT. Leveraging the conjugate symmetry of FFT can result in better performance and efficient model complexity. Instead of static weight in conventional circular convolution, this work dynamically produces complex weight matrices of circular convolution conditioned on the input features. And this operator also mixes the information of a complex weight tensor along the channel dimension. This channel mixing can complement circular convolution that is separable and input-independent. Experiments are conducted on both large and small datasets, and the DCCNet achieves better performance than other methods. It verifies the effectiveness of the proposed methods and its generalization capability.

Acknowledgement

This result was supported by "Region Innovation Strategy (RIS)" through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE)(2021RIS-003).



(a) Amplitude Spectrum (b) Phase Spectrum

Fig. 3. The amplitude spectrum (a) and phase spectrum (b) of the dynamic circular convolution.

References

1. Chi, L., Jiang, B., Mu, Y.: Fast fourier convolution. Advances in Neural Information Processing Systems **33**, 4479–4488 (2020)
2. Chu, X., Tian, Z., Zhang, B., Wang, X., Wei, X., Xia, H., Shen, C.: Conditional positional encodings for vision transformers. arXiv preprint arXiv:2102.10882 (2021)
3. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
4. Dai, Z., Liu, H., Le, Q.V., Tan, M.: Coatnet: Marrying convolution and attention for all data sizes. Advances in Neural Information Processing Systems **34**, 3965–3977 (2021)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
6. Guibas, J., Mardani, M., Li, Z., Tao, A., Anandkumar, A., Catanzaro, B.: Efficient token mixing for transformers via adaptive fourier neural operators. In: International Conference on Learning Representations (2021)
7. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y.: Transformer in transformer. Advances in Neural Information Processing Systems **34**, 15908–15919 (2021)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Heo, B., Yun, S., Han, D., Chun, S., Choe, J., Oh, S.J.: Rethinking spatial dimensions of vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11936–11945 (2021)
10. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
11. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
12. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84–90 (2017)
14. Li, J., Xia, X., Li, W., Li, H., Wang, X., Xiao, X., Wang, R., Zheng, M., Pan, X.: Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios. arXiv preprint arXiv:2207.05501 (2022)
15. Li, Y., Yuan, G., Wen, Y., Hu, E., Evangelidis, G., Tulyakov, S., Wang, Y., Ren, J.: Efficientformer: Vision transformers at mobilenet speed. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022), https://openreview.net/forum?id=NXHXoYMLIG
16. Li, Y., Zhang, K., Cao, J., Timofte, R., Van Gool, L.: Localvit: Bringing locality to vision transformers. arXiv preprint arXiv:2104.05707 (2021)
17. Liu, H., Dai, Z., So, D., Le, Q.V.: Pay attention to mlps. Advances in Neural Information Processing Systems **34**, 9204–9215 (2021)

18. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12009–12019 (2022)
19. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
20. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
21. Mehta, S., Rastegari, M.: Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=vh-0sUt8HlG
22. Mehta, S., Rastegari, M.: Separable self-attention for mobile vision transformers. arXiv preprint arXiv:2206.02680 (2022)
23. Min, J., Zhao, Y., Luo, C., Cho, M.: Peripheral vision transformer. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022), https://openreview.net/forum?id=nE8IJLT7nW-
24. Rao, Y., Zhao, W., Tang, Y., Zhou, J., Lim, S.L., Lu, J.: Hornet: Efficient high-order spatial interactions with recursive gated convolutions. Advances in Neural Information Processing Systems (NeurIPS) (2022)
25. Rao, Y., Zhao, W., Zhu, Z., Lu, J., Zhou, J.: Global filter networks for image classification. Advances in Neural Information Processing Systems **34**, 980–993 (2021)
26. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
27. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
29. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2149–2159 (2022)
30. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
31. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
33. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 568–578 (2021)

34. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media **8**(3), 415–424 (2022)
35. Wightman, R.: Pytorch image models. https://github.com/rwightman/pytorch-image-models (2019). https://doi.org/10.5281/zenodo.4414861
36. Xu, Y., Zhang, Q., Zhang, J., Tao, D.: Vitae: Vision transformer advanced by exploring intrinsic inductive bias. Advances in Neural Information Processing Systems **34**, 28522–28535 (2021)
37. Yang, J., Li, C., Dai, X., Gao, J.: Focal modulation networks. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022), https://openreview.net/forum?id=ePhEbo039l
38. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10819–10829 (2022)
39. Yu, W., Si, C., Zhou, P., Luo, M., Zhou, Y., Feng, J., Yan, S., Wang, X.: Metaformer baselines for vision. arXiv preprint arXiv:2210.13452 (2022)
40. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 558–567 (2021)