FcaNet: Frequency Channel Attention Networks

Zequn Qin[1], Pengyi Zhang[1], Fei Wu[1,2], Xi Li[1,2*]
[1]College of Computer Science, Zhejiang University,
[2]Shanghai Institute for Advanced Study, Zhejiang University
zequnqin@gmail.com, pyzhang@zju.edu.cn, wufei@cs.zju.edu.cn, xilizju@zju.edu.cn

Abstract

Attention mechanism, especially channel attention, has gained great success in the computer vision field. Many works focus on how to design efficient channel attention mechanisms while ignoring a fundamental problem, i.e., channel attention mechanism uses scalar to represent channel, which is difficult due to massive information loss. In this work, we start from a different view and regard the channel representation problem as a compression process using frequency analysis. Based on the frequency analysis, we mathematically prove that the conventional global average pooling is a special case of the feature decomposition in the frequency domain. With the proof, we naturally generalize the compression of the channel attention mechanism in the frequency domain and propose our method with multi-spectral channel attention, termed as FcaNet. FcaNet is simple but effective. We can change a few lines of code in the calculation to implement our method within existing channel attention methods. Moreover, the proposed method achieves state-of-the-art results compared with other channel attention methods on image classification, object detection, and instance segmentation tasks. Our method could consistently outperform the baseline SENet, with the same number of parameters and the same computational cost. Our code and models are publicly available at https://github.com/cfzd/FcaNet.

1. Introduction

As an important and challenging problem in feature modeling, attention mechanisms for convolutional neural networks (CNNs) have recently attracted considerable attention and are widely used in many fields like computer vision [40] and natural language processing [34]. In principle, they aim at selectively concentrating on some important information and have many types of variants (e.g., spatial attention, channel attention, and self-attention) corresponding
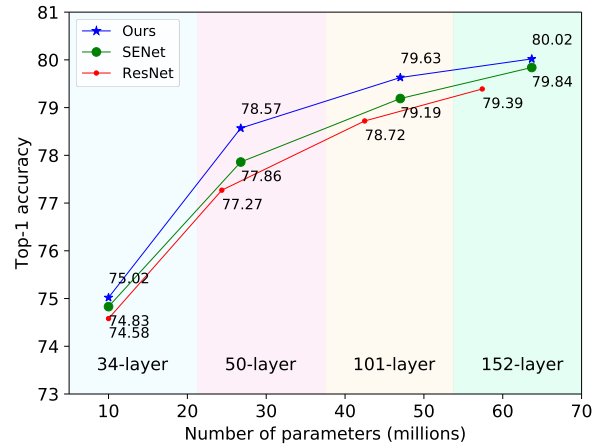
---

*Corresponding author.

Figure 1. Classification accuracy comparison on ImageNet. With the same number of parameters and computational cost, our method consistently outperforms the baseline SENet.

to different feature dimensions. Due to the simplicity and effectiveness in feature modeling, channel attention directly learns to attach importance weights with different channels, becoming a popular and powerful tool for the deep learning community.

Typically, a core step of channel attention approaches is to use a scalar for each channel to conduct the calculation due to the constrained computational overhead, and global average pooling (GAP) becomes the de-facto standard choice in the deep learning community because of its simplicity and efficiency. Nevertheless, every rose has its thorn. The simplicity of GAP makes it hard to well capture complex information for various inputs. Some methods like CBAM [39] and SRM [23] further use global max pooling and global standard deviation pooling to enhance the performance of GAP. Different from previous works, we consider the scalar representation of a channel as a compression problem. Namely, the information of a channel should be compactly encoded by a scalar while preserving the representation ability of the whole channel as much as possible. In this way, how to effectively compress a channel with a

scalar due to the constrained computational overhead is a major difficulty, and it is crucial to channel attention.

With the above motivation, we propose to use the discrete cosine transform (DCT) to compress channels in the channel attention mechanism for the following reasons: 1) DCT is a widely used data compression method in signal processing, especially with digital images and videos. Many widely used image and video formats like JPEG, HEIF, MPEG, and H.26x use DCT to realize data compression. DCT has a strong energy compaction property [1, 30], so it could achieve high data compression ratio with high quality [2, 22]. This property meets the demand of the channel attention that representing a channel with a scalar. 2) DCT can be implemented with an element-wise multiplication, and it is differentiable. In this way, it can be easily integrated into CNNs. 3) Surprisingly, DCT can be viewed as a generalization of GAP. Mathematically, GAP (showing the effectiveness in SENet [21]) is only equivalent to the lowest frequency components of DCT, leaving many other potentially useful frequency components unexplored. This strongly motivates us to tailor DCT for the channel attention mechanism.

In this paper, we further propose a simple, novel, but effective multi-spectral channel attention (MSCA) framework based on the above discussions. In order to better compress channels and explore the components left out by GAP, we propose to tailor DCT and use multiple but limited frequency components of DCT for the channel attention mechanism. Note that although we use multi-spectral channel attention, each channel is still represented by only one scalar. Along with the MSCA framework, how to select the frequency component of DCT for each channel is important. In this way, we propose three kinds of frequency component selection criteria to fulfill and validate the MSCA framework, which are LF (Low Frequency based selection), TS (Two-Step selection), and NAS (Neural Architecture Search selection). With these selection criteria, our method achieves state-of-the-art performance against the other channel attention ones.

In a word, the main contribution of this work can be summarized as follows.

- We regard the channel attention as a compression problem and introduce DCT in the channel attention. We then prove that conventional GAP is a special case of DCT. Based on this proof, we generalize the channel attention in the frequency domain and propose our method with the multi-spectral channel attention framework, termed as FcaNet.

- We propose three kinds of frequency component selection criteria along with the proposed multi-spectral channel attention framework to fulfill FcaNet.

- Extensive experiments demonstrate the proposed

method achieves state-of-the-art results on both ImageNet and COCO datasets, with the same computational cost as SENet. The results on ImageNet are shown in Fig. 1.

2. Related Work

Attention Mechanism in CNNs In [40], a visual attention method is first proposed to model the importance of features in the image caption task. Then many methods start to focus on the attention mechanism. A residual attention network [35] is proposed with a spatial attention mechanism using downsampling and upsampling. Besides, SENet [21] proposes the channel attention mechanism. It performs GAP on the channels and then calculates the weights of each channel using fully connected layers. What's more, GE [20] uses spatial attention to better exploit the feature context, and A^2-Net [7] builds a relation function for image or video recognition.

Inspired by these works, a series of works like BAM [28], DAN [9], CBAM [39], scSE [32], and CoordAttention [19] are proposed to fuse spatial attention [45] and channel attention. Among them, CBAM claims that GAP could only get a sub-optimal feature because of the loss of information. For addressing this problem, it uses both the GAP and the global max pooling and gains significant performance improvement. Similarly, SRM [23] also propose to use GAP with global standard deviation pooling. Motivated by CBAM, GSoP [11] introduces a second-order pooling method for downsampling. NonLocal [37] proposes to build a dense spatial feature map. AANet [3] proposes to embed the attention map with position information into the feature. SkNet [24] introduces a selective channel aggregation and attention mechanism, and ResNeSt [42] proposes a similar split attention method. Due to the complicated attention operation, these methods are relatively large. To improve efficiency, GCNet [4] proposes to use a simple spatial attention module and replace the original spatial downsampling process. ECANet [36] introduces one-dimensional convolution layers to reduce the redundancy of fully connected layers and obtains more efficient results.

Besides these works, many methods try to extend the attention mechanism to specific tasks, like multi-label classification [14], saliency detection [44], visual explanation [10], and super-resolution [43].

Frequency Domain Learning Frequency analysis has always been a powerful tool in the signal processing field. In recent years, some applications of introducing frequency analysis in the deep learning field emerge. In [8, 13], frequency analysis is introduced in the CNNs by JPEG encoding. Then, DCT is incorporated in [41] to reduce communication bandwidth. There are also some applications in the model compression and pruning tasks like [6, 27, 38].

3. Method

In this section, we first revisit the formulation of DCT and channel attention. Then, based on these works, we elaborate on the derivation of our multi-spectral channel attention framework. Meanwhile, along with the multi-spectral channel attention framework, three kinds of frequency components selection methods are proposed.

3.1. Revisiting DCT and Channel Attention

We first elaborate on the definitions of discrete cosine transform and channel attention mechanism.

Discrete Cosine Transform (DCT) Typically, the basis function of two-dimensional (2D) DCT [1] is:

$$B_{h,w}^{i,j} = \cos(\frac{\pi h}{H}(i + \frac{1}{2})) \cos(\frac{\pi w}{W}(j + \frac{1}{2})). \quad (1)$$

Then the 2D DCT can be written as:

$$f_{h,w}^{2d} = \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} x_{i,j}^{2d} B_{h,w}^{i,j} \quad (2)$$

$$s.t. \ \ h \in \{0, 1, \cdots, H-1\}, w \in \{0, 1, \cdots, W-1\},$$

in which $f^{2d} \in \mathbb{R}^{H \times W}$ is the 2D DCT frequency spectrum, $x^{2d} \in \mathbb{R}^{H \times W}$ is the input, H is the height of x^{2d}, and W is the width of x^{2d}. Correspondingly, the inverse 2D DCT can be written as:

$$x_{i,j}^{2d} = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} f_{h,w}^{2d} B_{h,w}^{i,j}, \quad (3)$$

$$s.t. \ \ i \in \{0, 1, \cdots, H-1\}, j \in \{0, 1, \cdots, W-1\}.$$

Please note that in Eqs. 2 and 3, some constant normalization factors are removed for simplicity, which will not affect the results in this work.

Channel Attention The channel attention mechanism is widely used in CNNs. It uses scalar to represent and evaluate the importance of each channel. Suppose $X \in \mathbb{R}^{C \times H \times W}$ is the image feature tensor in networks, C is the number of channels, H is the height of the feature, and W is the width of the feature. As discussed in Sec. 1, we treat the scalar representation in channel attention as a compression problem since it has to represent the whole channel while only one scalar can be used. In this way, the attention mechanism can be written as:

$$att = sigmoid(fc(compress(X))), \quad (4)$$

where $att \in \mathbb{R}^C$ is the attention vector, $sigmoid$ is the Sigmoid function, fc represents the mapping functions like fully connected layer or one-dimensional convolution, and

$compress : \mathbb{R}^{C \times H \times W} \mapsto \mathbb{R}^C$ is a compression method. After obtaining the attention vector of all C channels, each channel of input X is scaled by the corresponding attention value:

$$\widetilde{X}_{:,i,:,:} = att_i X_{:,i,:,:}, \quad s.t. \ \ i \in \{0, 1, \cdots, C-1\}, \quad (5)$$

in which $\widetilde{X}$ is the output of attention mechanism, att_i is the i-th element of attention vector, and $X_{:,i,:,:}$ is the i-th channel of input.

Typically, global average pooling is the de-facto compression method [21, 36] for its simplicity and effectiveness. There are also compression methods like global max pooling [39] and global standard deviation pooling [23].

3.2. Multi-Spectral Channel Attention

In this section, we first theoretically discuss the problem of existing channel attention mechanisms. Based on the theoretical analysis, we then elaborate on the network design of the proposed method.

Theoretical Analysis of Channel Attention As discussed in Sec. 3.1, DCT can be viewed as a weighted sum of inputs. We further prove that GAP is actually a special case of 2D DCT.

Theorem 1. *GAP is a special case of 2D DCT, and its result is proportional to the lowest frequency component of 2D DCT.*

Proof. Suppose h and w in Eq. 2 are 0, we have:

$$
\begin{aligned}
f_{0,0}^{2d} &= \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} x_{i,j}^{2d} \cos(\frac{0}{H}(i + \frac{1}{2})) \cos(\frac{0}{W}(j + \frac{1}{2})) \\
&= \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} x_{i,j}^{2d} \\
&= gap(x^{2d})HW.
\end{aligned}
$$

$$(6)$$

In Eq. 6, $f_{0,0}^{2d}$ represents the lowest frequency component of 2D DCT, and it is proportional to GAP. In this way, Theorem 1 is proved. ∎

Multi-Spectral Channel Attention Module Based on the theoretical analysis and Theorem 1, we can see that using GAP in the channel attention mechanism means only the lowest frequency information is preserved. And all components from other frequencies are discarded, which also encode the useful information patterns in representing the channels and should not be left out.

To better compress channels and introduce more information, we propose to generalize GAP to more frequency

(a) Original SENet

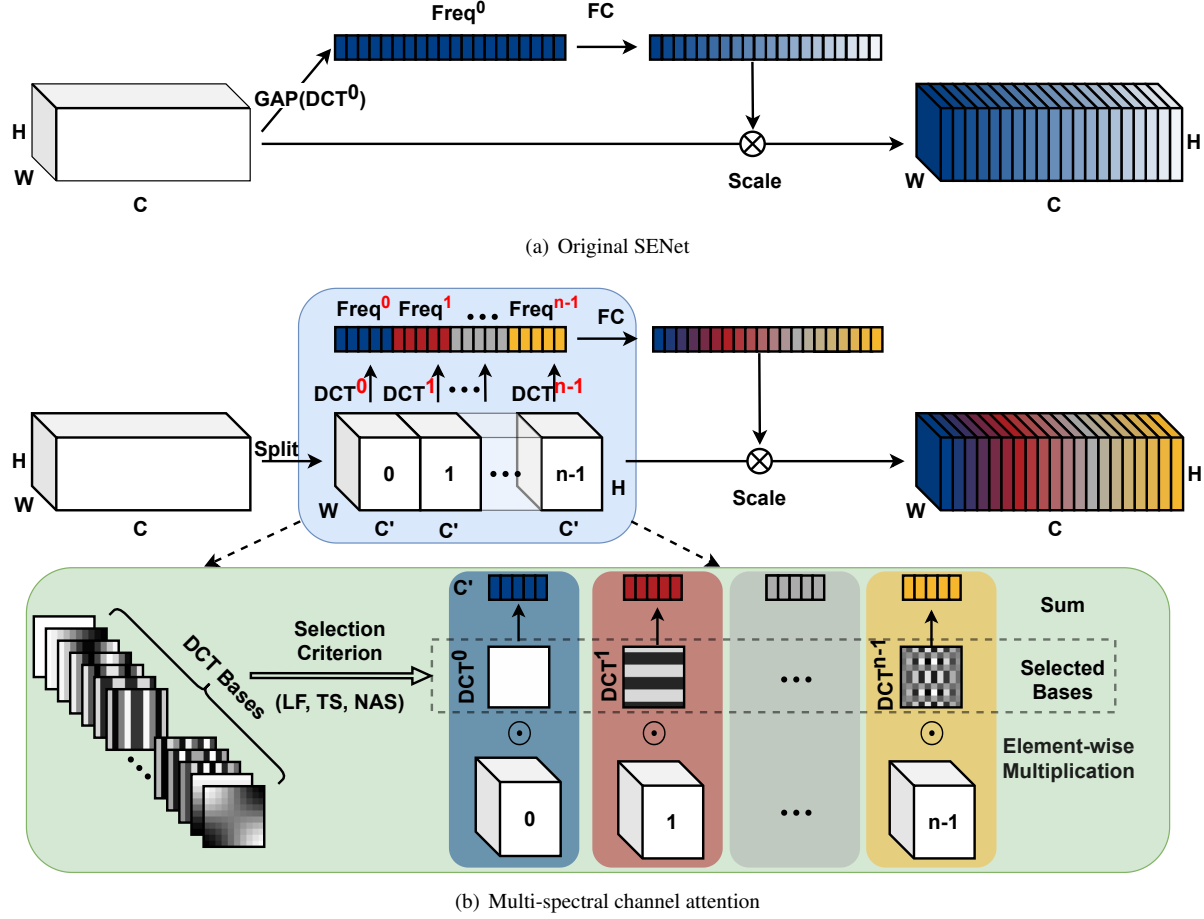(b) Multi-spectral channel attention

Figure 2. Illustration of existing channel attention and multi-spectral channel attention. For simplicity, the 2D DCT indices are represented in the one-dimensional format. We can see that our method uses multiple frequency components with the selected DCT bases, while SENet only uses GAP in channel attention. Best viewed in color.

components of 2D DCT and compress more information with multiple frequency components of 2D DCT, including the lowest frequency component, i.e., GAP.

First, the input X is split into many parts along the channel dimension. Denote $[X^0, X^1, \cdots, X^{n-1}]$ as the parts, in which $X^i \in \mathbb{R}^{C' \times H \times W}$, $i \in \{0, 1, \cdots, n-1\}$, $C' = \frac{C}{n}$, and C should be divisible by n. For each part, a corresponding 2D DCT frequency component is assigned, and the 2D DCT results can be used as the compression results of channel attention. In this way, we have:

$$Freq^i = 2\text{DDCT}^{u_i, v_i}(X^i),$$
$$= \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} X^i_{:,h,w} B^{u_i,v_i}_{h,w} \qquad (7)$$
$$s.t. \ i \in \{0, 1, \cdots, n-1\},$$

in which $[u_i, v_i]$ are the frequency component 2D indices corresponding to X^i, and $Freq^i \in \mathbb{R}^{C'}$ is the C'-dimensional vector after the compression. The whole com-

pression vector can be obtained by concatenation:

$$Freq = compress(X)$$
$$= \text{cat}([Freq^0, Freq^1, \cdots, Freq^{n-1}]), \qquad (8)$$

in which $Freq \in \mathbb{R}^C$ is the obtained multi-spectral vector. The whole multi-spectral channel attention framework can be written as:

$$ms\_att = sigmoid(fc(Freq)). \qquad (9)$$

From Eqs. 8 and 9, we can see that our method generalizes the original GAP method to a framework with multiple frequency components. By doing so, the channel information after compression is effectively enriched for representation. The overall illustration of our method is shown in Fig. 2.

Criteria for Choosing Frequency Components There exists an important problem of how to choose frequency

component indices $[u_i, v_i]$ for each part X^i. In order to fulfill the multi-spectral channel attention, we propose three kinds of criteria, which are FcaNet-LF (Low Frequency), FcaNet-TS (Two-Step selection), and FcaNet-NAS (Neural Architecture Search).

FcaNet-LF means FcaNet with low-frequency components. As we all know, many compression methods use low-frequency information of DCT to compress information. Moreover, some methods [21, 41] have shown CNNs prefer low-frequency information. In this way, the first criterion for choosing frequency components is to only select low-frequency components.

FcaNet-TS means FcaNet selects components within a two-step selection scheme. Its main idea is to first determine the importance of each frequency component and then investigate the effects of using different numbers of frequency components. Namely, we evaluate the results of each frequency component in channel attention individually. Finally, we choose the Top-k highest performance frequency components based on the evaluation results.

FcaNet-NAS means FcaNet with searched components. For this criterion, we use neural architecture search to search the best frequency component for channels. For each part X^i, a set of continuous variables $\alpha = \{\alpha^{(u,v)}\}$ are assigned to search components. The frequency components of this part can be written as:

$$Freq_{nas}^i = \sum_{(u,v)\in O} \frac{\exp(\alpha^{(u,v)})}{\sum_{(u',v')\in O}\exp(\alpha^{(u',v')})} 2\text{DDCT}^{u,v}(X^i),$$
(10)

in which O is the set containing all 2D DCT frequency component indices. After training, the frequency component for X^i is derived by $(u_i^*, v_i^*) = \arg\max_{(u,v)\in O}\{\alpha^{(u,v)}\}$.

The ablation studies about these criteria can be seen in Sec. 4.2.

4. Experiments

In this section, we first elaborate on the details of our experiments. Second, we show the ablation studies about FcaNet. Third, we give discussions about how the information is compressed in our framework, complexity, and code implementation. At last, we investigate the effectiveness of our method on the task of image classification, object detection, and instance segmentation.

4.1. Implementation Details

To evaluate the results of the proposed FcaNet on ImageNet [33], we employ four widely used CNNs as backbone models, including ResNet-34, ResNet-50, ResNet-101, and ResNet-152. We follow the data augmentation and hyperparameter settings in [16] and [17]. Concretely, the input images are cropped randomly to 224×224 with random horizontal flipping. We use an SGD optimizer with a momen-

tum of 0.9, a weight decay of 1e-4, and a batch size of 128 per GPU at training time. For large models like ResNet-101 and ResNet-152, the batch size is set to 64. The learning rate is set to 0.1 for a batch size of 256 with the linear scaling rule [12]. All models are trained within 100 epochs with cosine learning rate decay and label smoothing. Notably, for training efficiency, we use the Nvidia APEX mixed precision training toolkit.

To evaluate our method on MS COCO [26] using Faster R-CNN [31] and Mask R-CNN [15], we use the implementation of detectors from the MMDetection [5] toolkit and employ its default settings. During training, the shorter side of the input image is resized to 800. All models are optimized using SGD with a weight decay of 1e-4, a momentum of 0.9, and a batch size of 2 per GPU within 12 epochs. The learning rate is initialized to 0.01 and is decreased by the factor of 10 at the 8th and 11th epochs, respectively.

All models are implemented in PyTorch [29] framework and with eight Nvidia RTX 2080Ti GPUs.

4.2. Ablation Study

As discussed in Sec. 3.2, we propose three kinds of criteria, including FcaNet-LF (Low Frequency), FcaNet-TS (Two-Step selection), and FcaNet-NAS (Neural Architecture Search). In this section, we first show the ablations about these variants. Then we discuss the relation between FcaNet and fully learnable channel attention.

The effects of individual frequency components For FcaNet-TS, the first step is to determine the importance of each frequency component. To investigate the effects of different frequency components individually in channel attention, we only use one frequency component at a time. We divide the whole 2D DCT frequency space into 7×7 parts since the smallest feature map size is 7×7 on ImageNet. In this way, there are in total of 49 experiments. To speed up the experiments, we first train a standard ResNet-50 network for 100 epochs as the base model. Then we add channel attention to the base model with different frequency components to verify the effects. All added models are trained within 20 epochs with a similar optimization setting in Sec. 4.1, while the learning rate is set to 0.02.

As shown in Fig. 3, we can see that using lower frequency could have better performance, which is intuitive and verifies the success of SENet. This also verifies the conclusion [41] that deep networks prefer low-frequency information. Nevertheless, interestingly, we can see that nearly all frequency components (except the highest component) have very small gaps ($<= 0.5\%$ Top-1 accuracy) between the lowest one, i.e., vanilla channel attention with GAP. This shows that other frequency components can also cope well with the channel attention mechanism, and it is effective to generalize the channel attention in the frequency domain.

Figure 3. Top-1 accuracies on ImageNet using different frequency components in channel attention individually.

The effects of different numbers of frequency components For FcaNet-LF, we verify the results of using K lowest-frequency components. For FcaNet-TS, we select Top-K highest performance frequency components in Fig. 3. For simplicity, K could be 1, 2, 4, 8, 16, or 32.
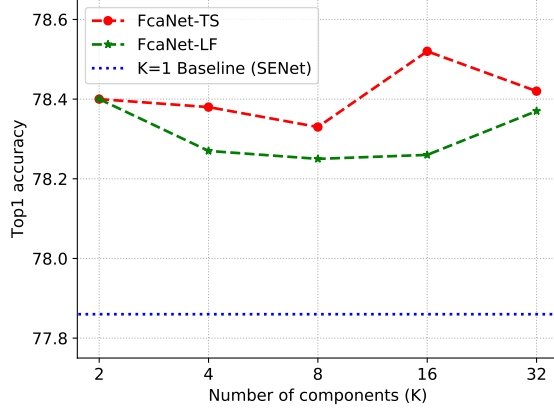

Figure 4. Top1 accuracy with different numbers of components. Since FcaNet-NAS searches and determines frequency components automatically, it is not included in this experiment.

As shown in Fig. 4, we can see two phenomena. 1) All experiments with multi-spectral attention have a significant performance gain compared with the one only using the GAP in channel attention. This verifies our idea of using multiple frequency components in channel attention. 2) For FcaNet-LF and FcaNet-TS, the settings with 2 and 16 frequency components gain the best performance, respectively. In this way, we use these settings in our method and all other experiments.

Comparison with fully learnable channel attention As shown in Eq. 7, we use the 2D DCT basis functions to com-

press channels. The 2D DCT basis functions $B_{h,w}^{u_i,v_i}$ can be simply regarded as a tensor containing DCT coefficients. In this way, a natural question is that how about directly learning a tensor to compress channels. We compare our method with three different kinds of tensors, which are Fixed tensor with Random initialization (FR), Learned tensor with Random initialization (LR), and Learned tensor with DCT initialization (LD). In this case, our method can be viewed as a Fixed tensor with DCT initialization (FD).
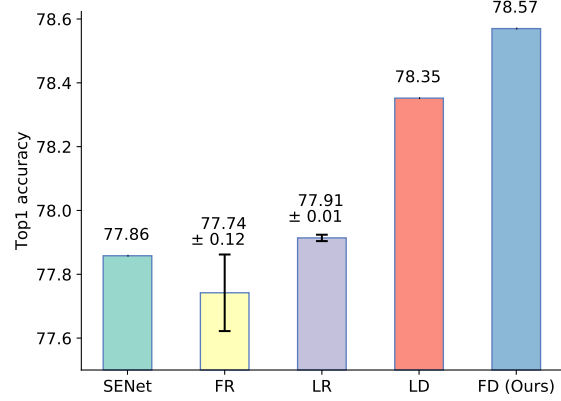

Figure 5. Comparison with fully learnable channel attention. FR means a Fixed tensor with Random initialization, LR means a Learned tensor with Random initialization, LD means a Learned tensor with DCT initialization, and FD means a Fixed tensor with DCT initialization, which is our method. For settings with the random initialization, the error bars are shown.

The comparison is shown in Fig. 5. We can see that all settings with the DCT initialization (LD, FD) outperform the ones without DCT (FR, LR). Moreover, the setting with fixed DCT initialization (ours) even outperforms the fully learnable channel attention methods (LR, LD), which shows the effectiveness of using DCT to compress channels.

4.3. Discussion

How the multi-spectral framework compresses and embeds more information In Sec. 3.2, we show that only using GAP in channel attention is actually discarding information of all other frequency components except the lowest one, i.e., GAP. In this way, generalizing channel attention in the frequency domain and using the multi-spectral framework could naturally embed more information in the channel attention mechanism.

Besides the above derivation, we also give a thought experiment to show that more information could be embedded. As we all know, deep networks are redundant [18, 46]. If two channels are redundant for each other, we can only get the same information using GAP. However, in our multi-spectral framework, it is possible to extract more information from redundant channels because different frequency components contain different information. In this way, the

Table 1. Comparison of different attention methods on ImageNet. All results are reproduced and trained with the same training setting except AANet, which has no official code.

Method	Years	Backbone	Parameters	FLOPS	Train FPS	Test FPS	Top-1 acc	Top-5 acc
ResNet [16]	CVPR16		21.80 M	3.68 G	2898	3840	74.58	92.05
SENet [21]	CVPR18		21.95 M	3.68 G	2729	3489	74.83	92.23
ECANet [36]	CVPR20	ResNet-34	21.80 M	3.68 G	2703	3682	74.65	92.21
FcaNet-LF			21.95 M	3.68 G	2717	3356	**74.95**	92.16
FcaNet-TS			21.95 M	3.68 G	2717	3356	**75.02**	92.07
FcaNet-NAS			21.95 M	3.68 G	2717	3356	**74.97**	92.34
ResNet [16]	CVPR16		25.56 M	4.12 G	1644	3622	77.27	93.52
SENet [21]	CVPR18		28.07 M	4.13 G	1457	3417	77.86	93.87
CBAM [39]	ECCV18		28.07 M	4.14 G	1132	3319	78.24	93.81
GSoPNet1*[11]	CVPR19		28.29 M	6.41 G	1095	3029	79.01	94.35
GCNet [4]	ICCVW19	ResNet-50	28.11 M	4.13 G	1477	3315	77.70	93.66
AANet [3]	ICCV19		25.80 M	4.15 G	-	-	77.70	93.80
ECANet [36]	CVPR20		25.56 M	4.13 G	1468	3435	77.99	93.85
FcaNet-LF			28.07 M	4.13 G	1430	3331	**78.43**	94.15
FcaNet-TS			28.07 M	4.13 G	1430	3331	**78.57**	94.10
FcaNet-NAS			28.07 M	4.13 G	1430	3331	**78.46**	94.09
ResNet [16]	CVPR16		44.55 M	7.85 G	816	3187	78.72	94.30
SENet [21]	CVPR18		49.29 M	7.86 G	716	2944	79.19	94.50
AANet [3]	ICCV19		45.40 M	8.05 G	-	-	78.70	94.40
ECANet [36]	CVPR20	ResNet-101	44.55 M	7.86 G	721	3000	79.09	94.38
FcaNet-LF			49.29 M	7.86 G	705	2936	**79.46**	94.60
FcaNet-TS			49.29 M	7.86 G	705	2936	**79.63**	94.63
FcaNet-NAS			49.29 M	7.86 G	705	2936	**79.53**	94.64
ResNet [16]	CVPR16		60.19 M	11.58 G	559	2721	79.39	94.74
SENet [21]	CVPR18		66.77 M	11.60 G	508	2566	79.84	94.82
AANet [3]	ICCV19		61.60 M	11.90 G	-	-	79.10	94.60
ECANet [36]	CVPR20	ResNet-152	60.19 M	11.59 G	515	2619	79.86	94.80
FcaNet-LF			66.77 M	11.60 G	502	2387	**80.13**	94.90
FcaNet-TS			66.77 M	11.60 G	502	2387	**80.02**	94.89
FcaNet-NAS			66.77 M	11.60 G	502	2387	**79.96**	94.94

* Please note that although GSoPNet1 has higher performance, the computational cost is 1.5 times larger than ours.

proposed multi-spectral framework could embed more information in the channel attention mechanism.

Complexity analysis We analyze the complexity of our method from two aspects: the number of parameters and the computational cost. For the number of parameters, our method has no extra parameters compared with the baseline because the weights of DCT are predefined constants.

For the computational cost, our method has a negligible extra cost and can be viewed as having the same computational cost as SENet. With ResNet-34, ResNet-50, ResNet-101, and ResNet-152 backbones, the relative computational cost increases of our method are 0.04%, 0.13%, 0.11%, and 0.11% compared with SENet, respectively. More results can be found in Table 1.

A Few lines of code change Another important property of the proposed multi-spectral framework is that it can be easily carried out with existing channel attention implemen-

tations. The only difference between our method and SENet is the channel compression method (GAP vs. multi-spectral 2D DCT). As described in Sec. 3.1 and Eq. 7, 2D DCT can be viewed as a weighted sum of inputs. It can be simply implemented by element-wise multiplication and summation. In this way, our method could be easily integrated into arbitrary channel attention methods.

4.4. Image Classification on ImageNet

We compare our FcaNet with the state-of-the-art methods using ResNet-34, ResNet-50, ResNet-101, and ResNet-152 backbones on ImageNet, including SENet [21], CBAM [39], GSoP-Net1 [11], GCNet [4], AANet [3], and ECANet [36]. The evaluation metrics include both efficiency (i.e., network parameters, floating point operations per second (FLOPs), and frame per second (FPS)) and effectiveness (i.e., Top-1/Top-5 accuracy). As shown in Table 1, our method achieves the best performance nearly in all experimental settings.

Table 2. Object detection results of different methods on COCO val 2017.

Method	Detector	Parameters	FLOPs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-50		41.53 M	215.51 G	36.4	58.2	39.2	21.8	40.0	46.2
SENet		44.02 M	215.63 G	37.7	60.1	40.9	22.9	41.9	48.2
ECANet	Faster-RCNN	41.53 M	215.63 G	38.0	60.6	40.9	23.4	42.1	48.0
FcaNet-LF		44.02 M	215.63 G	**39.0**	61.3	41.9	23.4	42.5	49.7
FcaNet-TS		44.02 M	215.63 G	**39.0**	61.1	42.3	23.7	42.8	49.6
FcaNet-NAS		44.02 M	215.63 G	**39.0**	60.9	42.3	23.0	42.9	49.9
ResNet-101		60.52 M	295.39 G	38.7	60.6	41.9	22.7	43.2	50.4
SENet		65.24 M	295.58 G	39.6	62.0	43.1	23.7	44.0	51.4
ECANet	Faster-RCNN	60.52 M	295.58 G	40.3	62.9	44.0	24.5	44.7	51.3
FcaNet-LF		65.24 M	295.58 G	**41.3**	63.4	44.9	24.6	45.6	53.6
FcaNet-TS		65.24 M	295.58 G	**41.2**	63.3	44.6	23.8	45.2	53.1
FcaNet-NAS		65.24 M	295.58 G	**41.2**	63.3	44.9	24.7	45.2	53.0
ResNet-50		44.17 M	261.81 G	37.2	58.9	40.3	22.2	40.7	48.0
SENet		46.66 M	261.93 G	38.7	60.9	42.1	23.4	42.7	50.0
GCNet		46.69 M	261.94 G	39.4	61.6	42.4	-	-	-
ECANet	Mask-RCNN	44.17 M	261.93 G	39.0	61.3	42.1	24.2	42.8	49.9
FcaNet-LF		46.66 M	261.93 G	**40.3**	61.9	43.9	24.9	43.6	52.2
FcaNet-TS		46.66 M	261.93 G	**40.3**	62.0	44.1	25.2	43.9	52.0
FcaNet-NAS		46.66 M	261.93 G	**40.3**	61.9	43.9	24.9	43.6	52.2

4.5. Object Detection on MS COCO

Besides the classification task on ImageNet, we also evaluate our method on object detection task to verify its effectiveness and generalization ability. We use our FcaNet with FPN [25] as the backbone (ResNet-50 and ResNet-101) of Faster R-CNN and Mask R-CNN and test their performance on the MS COCO dataset. SENet, CBAM, GC-Net, and ECANet are used for comparison.

As shown in Table 2, our method could also achieve the best performance with both Faster-RCNN and Mask-RCNN framework. Identical to the classification task on ImageNet, FcaNet could also outperform SENet by a large margin with the same number of parameters and computational cost. Compared with the SOTA method ECANet, FcaNet could outperform it by 0.9-1.3% in terms of AP.

4.6. Instance Segmentation on MS COCO

Besides the object detection, we then test our method on the instance segmentation task. As shown in Table 3, our method outperforms other methods by a more considerable margin. Specifically, FcaNet outperforms GCNet by 0.5% AP, while the gaps between other methods are roughly 0.1-0.2%. These results verify the effectiveness of our method.

5. Conclusion

In this paper, we study a fundamental problem of channel attention, that is, how to represent channels and regard this problem as a compression process. We have proved that GAP is a special case of DCT and proposed the FcaNet with the multi-spectral attention module, which generalizes

Table 3. Instance segmentation results of different methods using Mask R-CNN on COCO val 2017.

Method	AP	AP_{50}	AP_{75}
ResNet-50	34.1	55.5	36.2
SENet	35.4	57.4	37.8
GCNet	35.7	58.4	37.6
ECANet	35.6	58.1	37.7
FcaNet-LF	**36.3**	58.3	38.6
FcaNet-TS	**36.2**	58.6	38.1
FcaNet-NAS	**36.3**	58.3	38.6

the existing channel attention mechanism in the frequency domain. Meanwhile, we have explored different combinations of frequency components in our multi-spectral framework and proposed three criteria for frequency components selection. With the same number of parameters and computational cost, our method could consistently outperform SENet. We also have achieved state-of-the-art performance on image classification, object detection, and instance segmentation compared with other channel attention methods. Moreover, FcaNet is simple yet effective. Our method could be implemented with only a few lines of code change.

Acknowledgements

References

[1] Nasir Ahmed, T Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE Trans. Comput.*, 100(1):90–93, 1974. 2, 3

[2] Mario Barbero, H Hofmann, and ND Wells. Dct source coding and current implementations for hdtv. *EBU Technical Review*, (251):22–33, 1992. 2

[3] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Int. Conf. Comput. Vis.*, pages 3286–3295, 2019. 2, 7

[4] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *IEEE Conf. Comput. Vis. Worksh.*, pages 0–0, 2019. 2, 7

[5] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5

[6] Wenlin Chen, James Wilson, Stephen Tyree, Kilian Q Weinberger, and Yixin Chen. Compressing convolutional neural networks in the frequency domain. In *SIGKDD*, pages 1475–1484, 2016. 2

[7] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Aˆ 2-nets: Double attention networks. In *Adv. Neural Inform. Process. Syst.*, pages 352–361, 2018. 2

[8] Max Ehrlich and Larry S Davis. Deep residual learning in the jpeg transform domain. In *Int. Conf. Comput. Vis.*, pages 3484–3493, 2019. 2

[9] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3146–3154, 2019. 2

[10] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10705–10714, 2019. 2

[11] Zilin Gao, Jiangtao Xie, Qilong Wang, and Peihua Li. Global second-order pooling convolutional networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3024–3033, 2019. 2, 7

[12] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 5

[13] Lionel Gueguen, Alex Sergeev, Ben Kadlec, Rosanne Liu, and Jason Yosinski. Faster neural networks straight from jpeg. In *Adv. Neural Inform. Process. Syst.*, pages 3933–3944, 2018. 2

[14] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 729–739, 2019. 2

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Int. Conf. Comput. Vis.*, pages 2961–2969, 2017. 5

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 5, 7

[17] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 558–567, 2019. 5

[18] Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Int. Conf. Comput. Vis.*, pages 1389–1397, 2017. 6

[19] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. *arXiv preprint arXiv:2103.02907*, 2021. 2

[20] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In *Adv. Neural Inform. Process. Syst.*, pages 9401–9411, 2018. 2

[21] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7132–7141, 2018. 2, 3, 5, 7

[22] William Lea. *Video on demand*. House of Commons Library, 1994. 2

[23] HyunJae Lee, Hyo-Eun Kim, and Hyeonseob Nam. Srm: A style-based recalibration module for convolutional neural networks. In *Int. Conf. Comput. Vis.*, pages 1854–1862, 2019. 1, 2, 3

[24] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 510–519, 2019. 2

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2117–2125, 2017. 8

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755, 2014. 5

[27] Zhenhua Liu, Jizheng Xu, Xiulian Peng, and Ruiqin Xiong. Frequency-domain dynamic pruning for convolutional neural networks. In *Adv. Neural Inform. Process. Syst.*, pages 1043–1053, 2018. 2

[28] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018. 2

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inform. Process. Syst.*, pages 8026–8037, 2019. 5

[30] K Ramamohan Rao and Ping Yip. *Discrete cosine transform: algorithms, advantages, applications*. Academic press, 2014. 2

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Adv. Neural Inform. Process. Syst.*, pages 91–99, 2015. 5

[32] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks. *IEEE Medical Imaging*, pages 540–549, 2018. 2

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, pages 211–252, 2015. 5

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, pages 5998–6008, 2017. 1

[35] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3156–3164, 2017. 2

[36] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11534–11542, 2020. 2, 3, 7

[37] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7794–7803, 2018. 2

[38] Yunhe Wang, Chang Xu, Shan You, Dacheng Tao, and Chao Xu. Cnnpack: Packing convolutional neural networks in the frequency domain. In *Adv. Neural Inform. Process. Syst.*, pages 253–261, 2016. 2

[39] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Eur. Conf. Comput. Vis.*, pages 3–19, 2018. 1, 2, 3, 7

[40] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Int. Conf. Mach. Learn.*, pages 2048–2057, 2015. 1, 2

[41] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1740–1749, 2020. 2, 5

[42] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020. 2

[43] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Eur. Conf. Comput. Vis.*, pages 286–301, 2018. 2

[44] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3085–3094, 2019. 2

[45] Xizhou Zhu, Dazhi Cheng, Zheng Zhang, Stephen Lin, and Jifeng Dai. An empirical study of spatial attention mechanisms in deep networks. In *Int. Conf. Comput. Vis.*, pages 6688–6697, 2019. 2

[46] Zhuangwei Zhuang, Mingkui Tan, Bohan Zhuang, Jing Liu, Yong Guo, Qingyao Wu, Junzhou Huang, and Jinhui Zhu. Discrimination-aware channel pruning for deep neural networks. In *Adv. Neural Inform. Process. Syst.*, pages 875–886, 2018. 6