New Generation Computing
https://doi.org/10.1007/s00354-024-00252-7

RESEARCH PAPER

MANet: Mixed Attention Network for Visual Explanation

Jingjing Bai[1] · **Yoshinobu Kawahara**[2,3]

Received: 31 August 2023 / Accepted: 26 February 2024

Abstract
Various visual explanation methods, such as CAM and Grad-CAM, have been proposed to visualize and interpret predictions made by CNNs. Recent efforts go beyond mere visual interpretability, aiming to enhance CNN performance through the utilization of these generated visual explanations. In this work, we propose MANet (Mixed Attention Network)—a network architecture that advances the stability of visual explanations through an adaptive feature refinement mechanism via a mixed attention module. Concurrently, the generated attention maps are harnessed to bolster network performance in image recognition tasks. Experimental findings underscore the efficacy of MANet, demonstrating improved visual stability and consistency. The proposed architecture not only surpasses baseline models in image classification and object detection tasks but also establishes a novel paradigm for synergizing visual interpretability and network performance enhancement.

Keywords CNN · Attention mechanism · Mixed attention · Visual explanation

1 Introduction

Given the remarkable achievements of Convolutional Neural Network (CNN) models across diverse computer vision tasks [1–7], the problem of comprehending and interpreting CNNs has much attention recently. To address this, visual explanation methods, including CAM [8], Grad-CAM [9], Grad-CAM++ [10], and ABN [11], have been extensively adopted for a variety of recognition tasks for this purpose.

✉ Jingjing Bai
bai.jingjing.282@s.kyushu-u.ac.jp

Yoshinobu Kawahara
kawahara@ist.osaka-u.ac.jp

1 Faculty of Mathematics, Kyushu University, Fukuoka, Japan

2 Graduate School of Information Science and Technology, Osaka University, Suita, Japan

3 Center for Advanced Intelligence Project, RIKEN, Wako, Japan

Published online: 23 May 2024

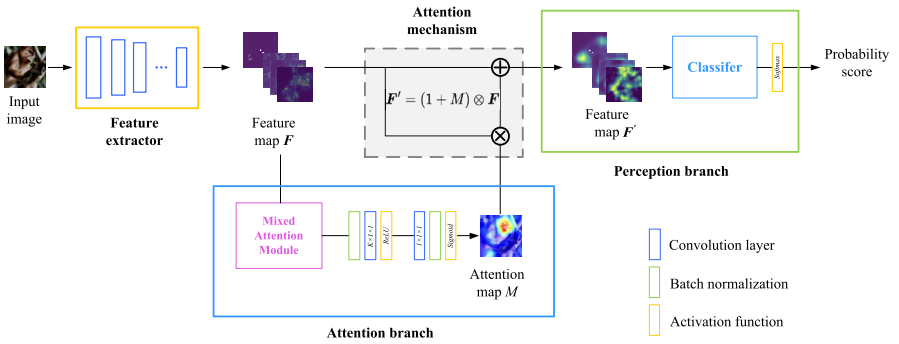

Fig. 1 Network structure of mixed attention network. The intermediate feature map is adaptively refined through the mixed attention module and generates a more consistent and stable attention map in the Attention Branch. $\oplus$ in the Attention Mechanism indicates element-wise summation, and $\otimes$ indicates element-wise multiplication

These aforementioned methods can be generally classified as either gradient-based explanations or response-based explanations.

Gradient-based methods, exemplified by methods like Grad-CAM [9] and Grad-CAM++ [10], excel in capturing the significance of specific regions within the input image for predictive output. However, as it requires extra backpropagation to derive the gradients from the output class to the input, it fails to provide a visual explanation in the forward pass. In contrast, response-based visual explanation methods like CAM [8] can generate the class activation maps (CAM) by projecting the weight matrix onto the channel-wise averaged feature maps. Notably, CAM replaces the conventional fully connected layer with Global Average Pooling (GAP) [12]. While this facilitates interpretability, it also introduces architecture sensitivity [9], potentially leading to model performance degradation [10, 11].

Inspired by response-based methods, Attention Branch Network (ABN) [11] is in some sense considered not only as a visual explanation method, but also improves the network performance through the incorporation of an attention-mechanism-based branch structure. Nevertheless, it has been observed in several experiments that ABN tends to produce unstable attention maps and hence leads to a decrease in classification accuracy. To tackle this concern, we present a network structure that extends the foundations of ABN, denoted as the Mixed Attention Network (MANet), which contains a mixed attention module that extracts informative features by combining cross-channel and spatial information, and then generates a more stable attention map using the "refined" feature during the inference process.

MANet, our novel network, which shares a structural resemblance with ABN, is composed of three parts: the feature extractor, the attention branch, and the perception branch, as shown in Fig. 1. The feature extractor is constructed with multiple convolutional layers, serving to extract feature maps from an input image. The attention branch, composed of a mixed attention module, adaptively refines the feature maps and generates a more reliable attention map, as illustrated in Fig. 2. Concurrently, the perception branch outputs the probability of a certain class through an attention mech-

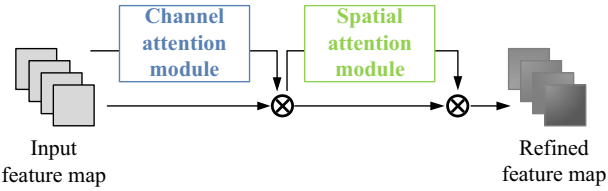

Fig. 2 Overview of mixed attention module. The module has two sequential sub-modules: *channel* and *spatial*

anism using the feature map from the feature extractor and the produced attention map.

Given that convolutional operations inherently capture informative information through the amalgamation of cross-channel and spatial features, our innovative mixed attention module is designed to emphasize significant information along these dimensions: the channel and spatial axes. This is accomplished by a sequential arrangement of channel and spatial attention modules, shown in Figs. 2 and 3. This sequential approach enables each submodule to discern 'what' and 'where' to allocate attention within the channel and spatial dimensions, respectively.

MANet can be effectively trained in an end-to-end manner and has demonstrated remarkable performance within image recognition tasks during our experiments. In the image classification experiment, we observed accuracy improvement from various existing networks using the ImageNet(1k) dataset. Additionally, we also validated the heightened performance of object detection task using the MS COCO dataset, thereby substantiating the efficacy of MANet. Lastly, we showed some illustrative examples of the attention maps generated by CAM, Grad-CAM, ABN, and our proposed method MANet, revealing MANet's propensity to allocate attention more accurately towards target objects compared to other methods. Additionally, the stability evaluation test shows that MANet always generates more stable and consistent attention maps than those produced by ABN.

Our contributions can be outlined as follows:

1. Innovative Architecture.
   We propose a novel architecture, known as MANet, for image recognition and visual explanation by integrating a mixed attention module for adaptive feature refinement. The incorporation of this mixed attention module sets MANet apart from existing models, enhancing its capabilities in both accurate recognition and effective visual explanation.
2. Validation of Mixed Attention Module.
   We thoroughly validate the efficacy of our proposed mixed attention module through comprehensive ablation studies. These experiments systematically demonstrate the impact of various architectural choices and design elements on the overall performance of MANet. By iteratively evaluating different configurations, we provide empirical evidence of the module's effectiveness in enhancing network capabilities.
3. Simultaneous Enhancement of Performance and Explanation.
   Another noteworthy contribution of our work is the integration of the attention

mechanism within the MANet architecture. This strategic integration facilitates the dual objectives of enhancing network performance while concurrently generating visual explanations that are more reliable, coherent, and consistent. This simultaneous achievement is a distinctive feature of MANet, signifying its potential to cater to both accuracy and interpretability demands within a single framework.

2 Related Work

2.1 Attention Branch Network

Attention Branch Network (ABN) [11] was introduced not only to enhance network performance, but also to concurrently generate better visual explanations within the forward pass. The ABN architecture consists of three components: the feature extractor, the attention branch, and the perception branch.

Within the attention branch of ABN, an overlapping 1×1 convolution scheme is employed to integrate and compress the feature map across the channel axes, and subsequently, the attention map is generated through a sigmoid function. Nonetheless, this process, while straightforward, proves to be rudimentary. As illustrated in the experiment below (Fig. 5), visual explanations generated in this way sometimes exhibit inconsistencies and instability. This can be attributed to the inherent limitations of the 1×1 convolution operation, which primarily reduces channel dimensions while inadequately extracting feature significance and contextual information along channel and spatial axes. Consequently, this inadequacy in coherent information extraction during the overlapping convolution process leads to a heightened degree of freedom, resulting in a decrease in network performance.

2.2 Attention Mechanism

The attention mechanism is a data processing technique widely employed in various machine learning tasks such as natural language processing [13–15], image recognition [16–19], and speech recognition [20–22]. The attention mechanism can help a model assign distinct weights to each part of input data and extract critical and relevant information to facilitate more accurate predictions. Additionally, it is achieved without significantly increasing computational or storage overheads, contributing to the widespread adoption of the attention mechanism. In recent years, most of the research works on the combination of deep learning and visual attention mechanisms have focused on the use of masks to form attention. The principle of the mask is to identify the key features in input image data through another layer of new weights. Through learning and training, a deep neural network with attention layers learns where to emphasize or suppress, and thus forms attention.

Certain notable works, such as [16, 17], employ spatial transformations on input data to extract essential information. These studies generate masks for spatial domains and assign weights. While spatial-wise attention focuses solely on spatial information and neglects channel-along information, treating input features within each channel

uniformly, it confines spatial domain transformation to the initial and shallow feature extraction phases, limiting the interpretability in deeper neural network layers. Conversely, channel-wise attention mechanisms, like [3], adaptively learn the importance of individual feature channels. However, their utilization of global average-pooled features for channel-wise attention computation somehow proves suboptimal for fine-grained channel attention inference. To address this, we advocate for the integration of max-pooled features as another effective alternative. Considering the remarkable performance of spatial-wise attention in directing focus, our MANet leverages both spatial and channel-wise attention by a novel mixed attention module. Through empirical validation in Sect. 4, we demonstrated the superior efficacy of exploiting both forms of attention rather than relying solely on one. Furthermore, our study includes a detailed exploration into the optimal configuration and arrangement of these two attention modules, with comprehensive details available in Sect. 4 for further elucidation.

3 Proposal Method

In this section, we introduce the network architecture of MANet. As previously indicated, MANet is composed of three components: the feature extractor, the attention branch, and the perception branch, as illustrated in Fig. 1.

The feature extractor comprises multiple convolution layers designed to extract informative feature maps from input data. Notably distinct from the attention branch of the ABN framework, MANet introduces a novel mixed attention module. This particular module, situated within the attention branch, serves a dual role in enhancing both the network's comprehension of feature significance and the stability of attention maps.

The mixed attention module enables the network to acquire a thorough understanding of the order of importance given to individual feature maps through the efficient manipulation of channel and spatial dimensions. This combination of mixed attention empowers the network with the capacity to comprehensively evaluate the intrinsic relevance of each feature map to the final decision-making process. Consequently, the generated attention maps exhibit an elevated level of consistency and reliability, enhancing the network's interpretability and predictive accuracy.

The perception branch predicts a class by exploiting the attention mechanism implemented upon the feature maps derived from the feature extractor, as well as the adaptively generated attention map. This collaboration results in a comprehensive decision-making process that takes advantage of the network's ability to intelligently emphasize relevant information guided by attention indications.

The entire network architecture is built upon a baseline model like ResNet [2]. Furthermore, the attention branch appears as a lightweight module that can be integrated into the baseline model after a specific layer. This incorporation then divides the baseline model into two parts: the feature extractor, responsible for information abstraction, and the perception branch, tasked with making accurate predictions. The next sections will go into great detail about each of these branches, and especially provide an extensive explanation of the mixed attention module's internal functions.

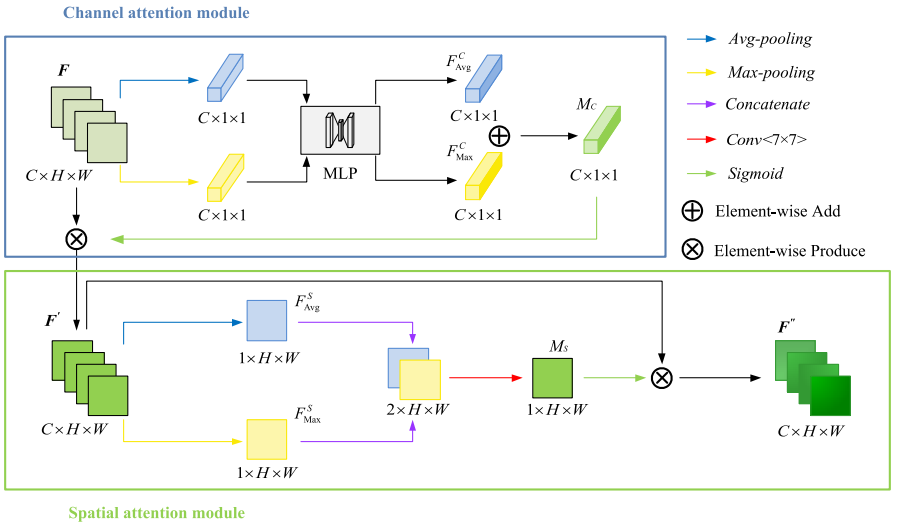

Fig. 3 Diagram of each attention sub-module. As illustrated, the channel-wise attention module utilizes both max-pooling outputs and average-pooling outputs with a shared MLP; the spatial-wise attention module utilizes similar two outputs that are pooled along the channel axis and forward them to a convolution layer

3.1 Mixed Attention Module

Provided a feature map $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$ as input to the mixed attention module, it sequentially computes a channel-wise attention map $M_c \in \mathbb{R}^{C\times 1\times 1}$ and a spatial-wise attention map $M_s \in \mathbb{R}^{1\times H\times W}$ at first, where $[C, H, W]$ format denotes the [channels, height, width] format of the provided feature map(as in color). Then the attention maps are multiplied by the input feature map for adaptive feature refinement. The entire process is illustrated in Fig. 3, which gives a visual overview of how the mixed attention module operates. In summary, the mixed attention module takes a feature map as input, calculates attention maps to emphasize important features, and then adjusts the original feature map accordingly for better feature representation, as summarized below:

$$\boldsymbol{F'} = M_c(\boldsymbol{F}) \otimes \boldsymbol{F},$$
$$\boldsymbol{F''} = M_s(\boldsymbol{F'}) \otimes \boldsymbol{F'},$$

where $\otimes$ indicates element-wise multiplication, $M_c(\boldsymbol{F})$ and $M_s(\boldsymbol{F'})$ respectively denote the corresponding attention maps calculated by each attention submodule given $\boldsymbol{F}$ and $\boldsymbol{F'}$ as input. As a result, $\boldsymbol{F''}$ is the final refined feature map computed by the mixed attention module.

This module is actually made up of two submodules working together. The first one, called the channel-wise attention module, focuses on identifying "what" is important by looking at each channel of the feature map. The second submodule, known as the spatial-wise attention module, concentrates on pinpointing "where" the significant parts are by examining different positions on the spatial dimensions. These two sub-

modules collaborate to help the network better understand both the crucial elements in each channel and the important positions in the spatial dimensions.

In the subsequent sections, we describe the specifics of each attention submodule, explaining how they work and how they contribute to the overall enhancement of the feature map.

3.1.1 Channel-Wise Attention Module

Since each channel in the feature map acts like a specialized detector for specific features [23], channel attention helps identify the "what" that holds significance when given an input image. Therefore, we generate a channel-wise attention map by examining the relationships between different channels of the feature map. To do this, we integrate and compress the spatial dimension of the input feature map.

Average pooling has been widely employed to aggregate spatial information. Besides, we suggest that max pooling, which highlights the most distinctive features of inputs, also provides important information for better channel attention. Thus, we combine the outputs of both average pooling and max pooling. We demonstrated empirically that combining both forms of pooled features results in significantly improved network representation capabilities as compared to employing them separately. (as elaborated in Sect. 4.1). Further details on how channel attention is computed are explained below.

We perform both average pooling and max pooling along the spatial axis of the feature map, placing them in a parallel manner to aggregate spatial information as optimally as possible, as illustrated in the upper part of Fig. 3. After the pooling operations, we obtain the average-pooled feature and the max-pooled feature. These features are then directed into a shared multi-layer perception (MLP) with one hidden layer. Note that the hidden activation size of MLP is defined as $\mathbb{R}^{C/r\times 1\times 1}$, where r signifies the reduction ratio. In our work, we set r to 16 based on insights gained from a comprehensive ablation study conducted by [3]. Consequently, we get the corresponding feature vectors F^c_{Avg} and F^c_{Max}. These feature vectors encapsulate valuable information about the importance of different features within each channel. Finally, the channel attention M_c is calculated by their element-wise sum. The entire procedure can be summarized as follows:

$$\begin{aligned} M_c &= \sigma(\text{MLP}(\text{AvgPool}(\boldsymbol{F})) + \text{MLP}(\text{MaxPool}(\boldsymbol{F}))) \\ &= \sigma(F^c_{\text{Avg}} + F^c_{\text{Max}}), \end{aligned}$$

where function AvgPool refers to average pooling operation and similarly function MaxPool refers to max pooling operation as described above. Function MLP refers to a shared multi-layer perception (MLP) with one hidden layer. σ indicates the sigmoid function. The attention channel M_c is then element-wise multiplied with the input feature F to get channel-wise refined feature map $\boldsymbol{F}'$, which is sequentially forwarded to the spatial attention module.

3.1.2 Spatial-Wise Attention Module

While channel attention is about identifying "what" features are significant, spatial attention is focused on figuring out "where" the informative parts are. Therefore, we construct a spatial attention map by leveraging the relationships between different spatial positions in the feature map. To do this, we integrate and compress the channel dimension of the input feature map this time.

We first perform average pooling and max pooling operations along the channel axis of the feature map. Then we concatenate the outputs of these pooling operations, F^s_{Avg} and F^s_{Max}, to form a single, more informative intermediate feature descriptor, as shown in the bottom half of Fig. 3. Finally, a convolution layer is applied to generate the spatial attention M_s which highlights informative regions on the input feature. The entire procedure can be summarized as follows:

$$\begin{aligned} M_s &= \sigma\left(\text{conv}^{7\times7}([\text{AvgPool}(\boldsymbol{F}); \text{MaxPool}(\boldsymbol{F})])\right) \\ &= \sigma\left(\text{conv}^{7\times7}([F^s_{\text{Avg}}; F^s_{\text{Max}}])\right), \end{aligned}$$

where σ indicates the sigmoid function, and function $\text{conv}^{7\times7}$ refers to convolution operation with the kernel size set to 7 [24], which is confirmed to be decent and effective by ablation study in Sect. 4.1.

In summary, we use the inter-spatial relationships between features to create a spatial attention map that tells the network where to pay attention. This complements the channel attention, creating a more comprehensive understanding of both "what" and "where" to focus on within the feature map.

3.1.3 Attention Branch

As the mixed attention module significantly contributes to the network's ability to identify what and where to emphasize or suppress, we get an adaptively refined feature $\boldsymbol{F''}$ that can localize the discriminative region more precisely. To generate an attention map in the training process, the attention branch constructs a top layer based on CAM [8], replacing the fully connected layer with a $K \times 1 \times 1$ convolution layer, as shown in Fig. 1. This operation yields a $K \times H \times W$ feature map, which in turn, is employed to create an initial attention map. To further consolidate the individual feature maps (K in total), they are collectively convoluted using a $1 \times 1 \times 1$ convolution layer, which results in the generation of a $1 \times H \times W$ feature map. To refine this feature map into a reliable attention map, we apply normalization using the sigmoid function and finally get the attention map $M \in \mathbb{R}^{1\times H\times W}$.

3.2 Perception Branch

The perception branch outputs the probability of a certain class using the feature maps from the feature extractor, which are subjected to an attention mechanism. As shown in Fig. 1, the attention map M generated from the attention branch is applied to the

intermediate feature map $\boldsymbol{F}$ via the following attention mechanism:

$$\boldsymbol{F}' = (1 + M) \otimes \boldsymbol{F}.$$

In this manner, the perception branch employs the attention map to guide the feature map toward identifying and emphasizing relevant features for accurate class prediction. The attention mechanism we used has a bypass structure so that it can efficiently improve the network performance by highlighting the feature map at the location with a higher value of the attention map while preventing the lower value region of the attention map from degrading to zero.

3.3 Training

The training process of MANet is conducted in an end-to-end manner, with the training loss calculated as a combination of the Softmax function and cross-entropy at the perception branch, specifically in the context of the image classification task. Importantly, the optimization of the mixed attention module is intrinsically associated with the attention mechanism for the perception branch. This incorporation ensures that the mixed attention module improves the network's classification performance without the need for an additional loss function.

Furthermore, MANet's adaptability extends beyond image classification. When applied to different image recognition tasks, the architecture demonstrates adaptability in terms of its training loss. This adaptability lies in the fact that the training loss can flexibly adapt to the underlying properties of the chosen baseline model. By dynamically adjusting the training loss according to the baseline model's attributes, MANet effectively optimizes itself for a wide range of image recognition tasks. This characteristic shows the flexibility and applicability of MANet across diverse domains of image analysis and interpretation.

In the forthcoming two sections, we hope to verify the efficacy and superiority of the proposed model through a two-stage experiment. Initially, detailed in Sect. 4, we verified the contributions and impacts of individual components within the mixed attention module described above and the optimal way of combining them through some ablation experiments. Subsequently, in Sect. 5, we presented a comparative analysis between MANet and established benchmarks across diverse image recognition tasks. This comparison serves to validate MANet's efficacy in improving network accuracy while substantiating the reliability and consistency of the visual explanations generated by MANet, fostering a comprehensive discussion on their stability.

4 Ablation Studies

In this section, we perform several ablation studies to demonstrate the effectiveness of our proposed design paradigm in mixed attention module. First, we present the experiment setup for our ablation studies. We use the ImageNet(1k) dataset for image classification task and ResNet50 as the baseline model. The dataset utilized in this

Table 1 Ablation study on the structure of the channel-wise attention module

	Top-1 error (%)	Top-5 error (%)
ResNet50 (baseline)	24.56	7.50
ResNet50 + AvgPool(SE [3])	23.14	6.70
ResNet50 + MaxPool	23.20	6.83
ResNet50 + AvgPool & MaxPool	**22.80**	**6.52**

We can observe that our proposed arrangement of the channel-wise attention module outperforms the SE [3] module
Bold values indicate that our model outperforms the other comparative models in this experiment

ablation study (also utilized in the following image classification experiment, see Sect. 5.1), namely ImageNet(1k), comprises a substantial training set of 1.2 million images and a validation subset (used as the testing dataset) encompassing 50,000 images. These images are annotated across a diverse spectrum of 1000 distinct object classes.

For the data augmentation, a two-step process is undertaken. Initially, the images are resized to dimensions of 256×256 pixels, following which they undergo a random cropping procedure to attain a final dimension of 224×224. The optimization of our network is executed via the stochastic gradient descent (SGD) algorithm. Commencing with an initial learning rate of 0.1, a decay is introduced at the culmination of every 30 epochs. The training regimen spans a total of 90 epochs, with a designated batch size of 256.

Since the mixed attention module is split into channel attention submodule and spatial attention submodule, here our process of designing modules is consequently structured into three phases: an exploration of an efficient methodology for computing channel-wise attention, followed by a similar investigation into spatial-wise attention and lastly, the combination of two submodules. See the subsequent subsections for the details.

4.1 Arrangement of Channel-Wise Attention

We incorporate both average pooling and max pooling within the channel-wise attention submodule to extract more refined feature information. To verify the rationale behind this configuration, we contrast three different approaches for computing channel attention: exclusive utilization of average pooling, exclusive utilization of max pooling, and the concurrent implementation of both pooling strategies, as initially conceptualized.

Table 1 presents the experimental results across the above pooling methodologies. We find that max-pooling just captures the informative information as average pooling compared to the baseline model. In addition, it is noteworthy that the channel attention module involving only average pooling aligns with the SE [3] module. Pertinently, the max-pooled feature encapsulates the most prominent information of the input feature map, whereas average pooling engenders a smoother feature representation that retains the essence of the features in the input. Consequently, the simultaneous employment of both pooling yields enhanced performance.

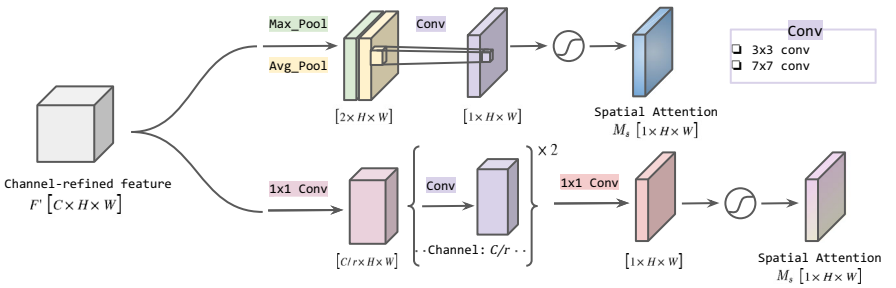

Fig. 4 Detail of the arrangement of spatial-wise attention module, where r is the same reduction ratio with the MLP in channel-wise attention module (see 3.1.1) for simplicity. Given the channel-wise refined feature, we conduct two different ways of extracting channel information and evaluate the influence of kernel size at the following convolution layer

Table 2 Ablation study on the structure of the spatial-wise attention module

	Top-1 error (%)	Top-5 error (%)
ResNet50 + Channel	22.80	6.52
ResNet50 + Channel + Spatial (avg & max, $k = 3$)	22.68	6.41
ResNet50 + Channel + Spatial (avg & max, $k = 7$)	**22.66**	**6.31**
ResNet50 + Channel + Spatial ($1 \times$ 1conv, $k = 3$)	22.96	6.64
ResNet50 + Channel + Spatial ($1 \times$ 1conv, $k = 7$)	22.90	6.47

We can observe that it performs better when employing the proposed average-and-max pooling along the channel axis, followed by the convolution operation with a larger kernel size of 7
Bold values indicate that our model outperforms the other comparative models in this experiment

4.2 Arrangement of Spatial-Wise Attention

As the spatial-wise attention module produces a 2D spatial attention map to emphasize or suppress features in different spatial locations, we adopt a symmetric methodology to compute the spatial-wise attention. We first apply channel-along average and max poolings, thereby establishing inter-spatial relationships among features. These pooled outputs are subsequently concatenated and convolved by a 7×7 convolution, ultimately obtaining the spatial attention map.

For the ablation study in the arrangement of spatial-wise attention module, we discuss two different ways to squeeze and extract channel information respectively: channel-along average-and-max pooling (see the top half of Fig. 4) and standard 1×1 convolution to integrate and compress the feature map across the channel axis (see the bottom half of Fig. 4). Furthermore, acknowledging the established significance of a broader receptive field in effectively leveraging contextual information and helping the network learn where the important spatial locations are, we extend our scrutiny to the subsequent convolutional layer, examining the impact of kernel sizes - specifically, those of 3×3 and 7×7.

Table 2 presents the experimental results across the above discussion. Note that in the spatial-wise attention ablation study, we position the spatial-wise attention

Table 3 Ablation study on the arrangement of two attention submodules

	Top-1 error (%)	Top-5 error (%)
ResNet50 (baseline)	24.56	7.50
ResNet50 + Channel + Spatial	**22.66**	**6.31**
ResNet50 + Spatial + Channel	22.78	6.42
ResNet50 + Channel & Spatial in parallel	22.95	6.59

We can observe that it performs best when using both two attention submodules and arranging them sequentially with the channel-wise attention going first
Bold values indicate that our model outperforms the other comparative models in this experiment

module subsequent to the earlier designed channel-wise attention module. For the comparison of different spatial attention approaches, it is shown that channel-along average-and-max pooling outperforms 1×1 convolution. For the comparison of different convolution kernel sizes, it can be observed that a larger kernel size provides higher accuracy in both cases, which supports the fact stated above: a larger receptive field implies greater certainty about discriminative spatial locations. What's more, from the experimental result, we can also find that choosing the suitable spatial attention module can improve the accuracy more than using only channel-wise attention module, which implies the necessity of applying both attention submodules. Putting it all together, we apply the channel-along average-and-max pooling followed by 7×7 convolutional layer as spatial-wise attention module.

4.3 Arrangement of Attention Modules

In this part of the ablation study, we evaluate sequential channel-spatial, sequential spatial-channel, and parallel implementation of both attention modules as three different arrangements of the channel and spatial attention submodules. Considering that different combinations may have an impact on the fineness of the final extracted features, we chose the above three arrangements.

As for the way of parallel implementation, since the two attention maps are different in size, we first expand them to $\mathbb{R}^{C \times H \times W}$ and then combine them by element-wise summation. Finally, a sigmoid function is used to compute the overall attention map and then element-wisely multiplied by the input feature map $\mathbf{F}$ to get the refined feature map $\mathbf{F}'$. The whole process can be described as:

$$\begin{aligned} \mathbf{F}' &= (1 + M) \otimes \mathbf{F} \\ &= (1 + \sigma(M_c \oplus M_s)) \otimes \mathbf{F}, \end{aligned}$$

where σ is a sigmoid function, $\oplus$ indicates element-wise summation, and the two attention maps are resized to $\mathbb{R}^{C \times H \times W}$ before element-wise summation.

Table 3 presents the experimental results across the above arrangements. It is shown that arranging the attention modules sequentially can produce higher accuracy than putting them in parallel. Especially, the sequential channel-spatial manner yields a slightly better performance compared to the reverse sequence.

Table 4 Comparison of top-1 errors on ImageNet(1k) dataset

	ResNet50	ResNet101	ResNet152
Original	24.56	22.38	22.24
SENet	23.14	22.35	21.57
ABN	23.14	22.11	21.82
MANet	**22.66**	**21.84**	21.95

We can observe that our MANet basically outperforms the previous works, especially when applying ResNet50 as the baseline model
Bold values indicate that our model outperforms the other comparative models in this experiment

4.4 Final Model Design

In a brief conclusion, we verified the structure and arrangement of the optimal channel-wise and spatial-wise attention modules through a series of ablation studies. Our final module is shown in Fig. 3. We use both average pooling and max pooling to extract spatial/channel information, and convolution with a kernel size of 7 in the spatial-wise attention module to obtain the contextual information. Finally, we arrange the two modules in a channel-spatial order, and experiments show that this design and combination greatly improve the accuracy compared to the baseline model and the related work SENet [3].

5 Experiments

In this section, we present several experimental results to evaluate the performance of our proposed MANet and show its effectiveness.

The following image recognition experiments are based on two standard datasets: ImageNet-1K for the image classification task(Sect. 5.1) and MS COCO for the object detection task(Sect. 5.2). To ensure a comprehensive and fair comparison, we took the initiative to reproduce all the networks [2, 3, 11] under examination within the PyTorch framework [25].

We also qualitatively evaluate our proposed MANet through visual explanation on ImageNet(1k) dataset(Sect. 5.3), and quantitatively validate the stability and consistency of those visualization generated by MANet(Sect. 5.3.1).

5.1 Image Classification

Firstly, we conducted image classification experiments on ImageNet(1k) dataset to thoroughly validate our model. We followed the same experimental setting described in Sect. 4 and verified our MANet in various network architectures. Specifically, we adopted the ResNet50, ResNet101, and ResNet152 as the baseline models for comparison. The experimental result of the top-1 errors on ImageNet(1k) is shown in Table 4.

We compare the accuracy of the original ResNet model, SENet [3], ABN [11], and MANet. As shown in the table, by introducing both channel and spatial attention maps, the top-1 errors of MANet on the ImageNet(1k) dataset are reduced by 1.9%, 0.54% and 0.29% compared to the baseline models. Moreover, compared to conventional models such as SENet or ABN, our proposed model reduces the top-1 error more significantly, especially when adopting ResNet50 as the baseline model, indicating the effectiveness of the mixed attention module.

However, while using ResNet152 as the baseline architecture, things become a little different. Although MANet has improved performance over the baseline model, it is still less accurate than either SENet or ABN. We speculate that this is caused because of the model complexity of ResNet152, which may lead to the non-convergence issue, making it difficult to achieve the expected accuracy.

5.2 Object Detection

Second, We conducted object detection on the Microsoft COCO dataset. This dataset involves 80 k training images (designated as "2014 train") and 40k validation images (labeled as "2014 val"). We used the training set as well as the subset of the validation set for training, holding out about 5000 images for testing. In quantifying the performance of our object detection experiment, we employed a well-accepted evaluation metric, mAP@.5, which is the mean average precision over $IoU_{\text{threshold}} = 0.5$. The equation is provided below:

$$\text{mAP} = \frac{\sum_{i=0}^{N-1} \int_0^1 P(R)\mathrm{d}R}{N};$$
$$R = \frac{\text{TP}}{\text{TP} + \text{FN}}; \quad P = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

where N denotes the total number of classes present within the dataset, R indicates the recall and P indicates the precision. Further delving into the equation, TP indicates the number of positive samples that have been correctly identified as positive, while FN indicates the number of positive samples that have been inaccurately identified as negative. On the other hand, FP indicates the number of negative samples that have been incorrectly predicted as positive. In the field of object detection, precision and recall are two important measurements used to evaluate the efficacy of detection algorithms. What's more, Average Precision, or AP, quantifies the area encompassed by the Precision-Recall (PR) curve in relation to the coordinate axes. A higher AP value is indicative of a more accurate target detection for a specific category.

Moreover, the notion of mean Average Precision (mAP) denotes the average value of the individual AP scores across various categories. A greater mAP value signifies a superior performance in object detection across the multitude of categories under consideration.

We adopted Faster-RCNN [6] as the detection method, which has proven to be robust and effective in object detection tasks. Furthermore, we utilized ImageNet pre-trained ResNet50 and ResNet101 as baseline models. The results of our experiments

Table 5 Object detection mAP@.5 on the MS COCO validation set

Backbone	Detector	mAP@.5
ResNet50	Faster-RCNN	46.2
MANet(ResNet50)	Faster-RCNN	**48.2**
ResNet101	Faster-RCNN	48.4
MANet(ResNet101)	Faster-RCNN	**50.5**

We adopted the Faster R-CNN [6] detection framework and implemented our model based on various baseline networks
Bold values indicate that our model outperforms the other comparative models in this experiment

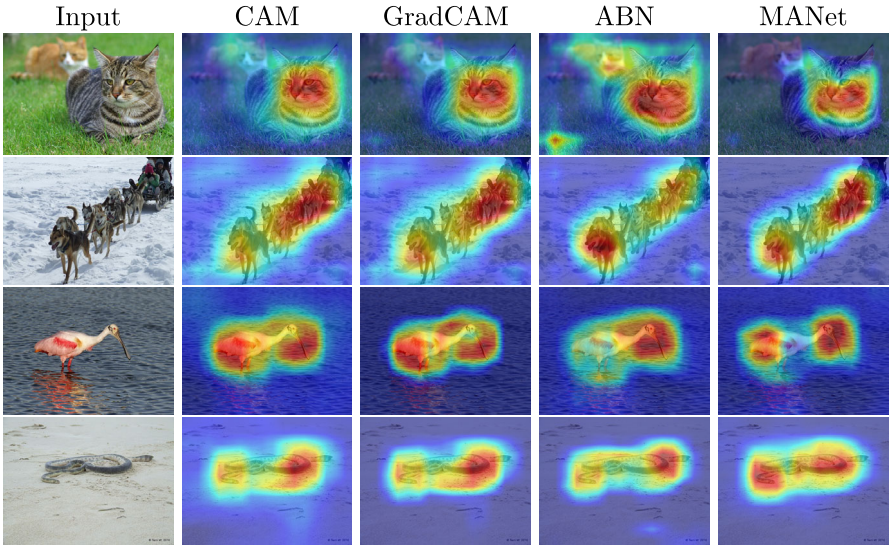

Fig. 5 Visual explanation results of various methods on ImageNet. Notably, the MANet always highlights the true class object correctly and in a more focused manner

are detailed in Table 5, where we showcase the outcomes of integrating MANet with the Faster R-CNN framework using the aforementioned baseline networks. Notably, our findings reveal a noteworthy improvement over the baseline models, indicative of the substantial performance enhancements brought forth by the incorporation of MANet. This empirical evidence attests to the efficacy and potential of our proposed approach in advancing object detection capabilities within the Faster R-CNN framework.

5.3 Visual Explanation

To comprehensively assess the quality of the generated visual explanations, we conducted a qualitative comparison among different visual explanation methods: CAM, Grad-CAM, ABN, and our proposed MANet. For this analysis, we utilized ResNet152 as the baseline model. Figure 5 shows the attention maps produced by each of these models using the ImageNet dataset.

While CAM, Grad-CAM, ABN, and MANet all exhibit a degree of similarity in terms of the regions they highlight, it is evident that MANet offers some notable advantages.

For a single object, the attention maps generated by MANet are tighter and focus on the salient features of the true class compared to any other methods. For example, in the case of the spoonbill in the third row, MANet distinctly emphasizes its defining characteristics, notably the long, paddle-shaped beak and its deep red wings and tail, acknowledged as pivotal and discriminative traits. Correspondingly, in the fourth row's sea snake, MANet also highlights the most discernable parts of the target object area in a less noisy and more focused manner, thereby enhancing its interpretability.

For multiple objects, MANet always shows better performance in locating the most distinguishable region of the target object. For instance, in the second row, the dog in the front of the image and the sled in the back are highlighted prominently because the model classifies this input image as "dogsled". As for the example in the first row, while ABN produces unreliable attention maps that inaccurately highlight unrelated grassy regions, MANet adeptly pinpoints the crucial features guiding its classification of the image as a 'tiger cat', specifically highlighting the striped cat occupying the main body of the image in front.

However, the notably constrained attention map produced by MANet in this instance prompts inquiries regarding unattended elements, such as the cat in the background, raising pertinent questions regarding holistic recognition. We address this phenomenon as "class discriminative visualization", albeit acknowledging potential limitations in its conclusive explanatory power. Notably, the MANet model is optimized not only by learning the feature importance but also by enhancing the backbone feature representation to focus more on important features to make decision for the input image. That is, in this case, by compressing and extracting the features, MANet focuses on primary influential features while disregarding secondary, albeit non-irrelevant ones. Consequently, the explanations derived from MANet remain equitable, task-aligned, and broadly compelling. Nevertheless, acknowledging the qualitative nature of our analysis, subjectivity and situational contexts may influence these outcomes. Therefore, a nuanced discussion and contextual analysis are recommended across various tasks and scenarios to provide comprehensive insights.

To conclude, this qualitative comparison substantiates that MANet provides enhanced performance and precision in generating visual explanations, thus enhancing its potential applicability in interpretability tasks within the domain of computer vision.

5.3.1 Stability Evaluation of Visual Explanation

In the previous section, we have shown several visual explanations generated by different models, and a simple comparison gave us an idea of how faithful the different visual explanations are to the detected target. Somehow, some tangible measurements are still preferred for real-world applicability. In this section, we try to provide a more convincing evaluation of the stability and consistency of these visual explanations. To do this, we introduce the following IoU-based measurement.

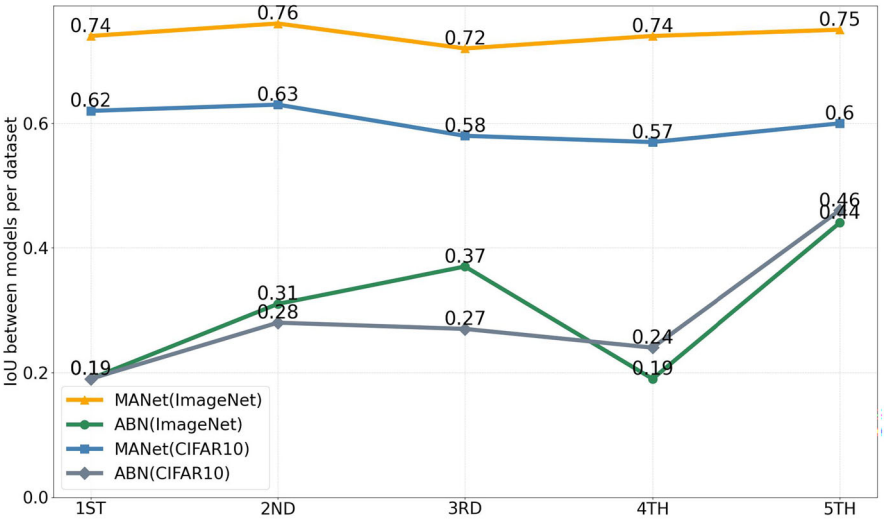

Fig. 6 Stability evaluation of visual explanation. We can observe that our MANet always generates stable and consistent visual explanations in contrast to previous work such as ABN

We first used ABN [11] as a comparison and did six independent training sessions for each of these models. Each session employed a different learning rate which is randomly sampled in the interval of [0.05, 0.11]. The baseline model used for stability evaluation is ResNet50 and the datasets on which we train the models individually and generate visual explanations are the ImageNet(1k) dataset and the CIFAR10 dataset. For the fully trained 6 models, we chose the one with the highest classification accuracy as the baseline model and calculated the IoU (Intersection of Union) between the visual explanations produced by the other five models with the baseline model. The threshold for computing the IoU value was empirically set to be ≥ 127 [10].

As shown in Fig. 6, our MANet shows great stability and consistency on both datasets. For the ImageNet(1k) dataset which consists of more object classes, MANet indicates a better high-attention-area coverage, as shown in Fig. 5. As for the stability of the generated visual explanation, MANet performs well on both datasets compared to ABN.

6 Conclusions

In this paper, we introduced the Mixed Attention Network(MANet), an innovative network architecture incorporating a mixed attention module, which is trainable in an end-to-end manner and able to generate improved visual explanations. Notably, the mixed attention module embedded within the attention branch significantly enhances the network's capacity to represent complex features, resulting in the production of more reliable attention maps through the utilization of mixed attention-based refined features.

To thoroughly evaluate the effectiveness of MANet, extensive experiments were conducted across image classification and object detection tasks. The experimental results demonstrated the superior performance of MANet when compared to various baseline models. Additionally, we also showed that the visualization explanations generated by MANet are more consistent, reliable, and in a more focused manner.

Our future research endeavors will encompass extending the application of MANet to diverse tasks, including fine-grained recognition, semantic segmentation, and multi-task learning. These upcoming explorations are aimed at further demonstrating the versatility and potency of MANet within the realm of computer vision, and subsequently contributing to the advancement of interpretability and performance in various real-world applications.

Funding Open Access funding provided by Osaka University.

Data availability All relevant data are within the paper.

Declarations

Conflict of interest There is no financial relationship to the works of this manuscript.

References

1. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
2. 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)
3. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
5. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
6. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. **28** (2015)
7. 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)
8. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)

9. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
10. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847 (2018)
11. Fukui, H., Hirakawa, T., Yamashita, T., Fujiyoshi, H.: Attention branch network: Learning of attention mechanism for visual explanation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10705–10714 (2019)
12. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
13. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2014)
14. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding (2018)
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
16. Huang, Z., Wang, X., Wei, Y., Huang, L., Shi, H., Liu, W., Huang, T.S.: CCNet: Criss-cross attention for semantic segmentation (2020)
17. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification (2017)
18. Luong, M.-T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
19. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
20. Liu, Z., Wang, L., Wu, W., Qian, C., Lu, T.: TAM: temporal adaptive module for video recognition (2021)
21. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 21–29 (2016)
22. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4651–4659 (2016)
23. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, pp. 818–833. Springer (2014)
24. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
25. Pytorch. http://pytorch.org/ Accessed: (2017-11-08)