EASR

Engineering and Applied Science Research

https://www.tci-thaijo.org/index.php/easr/index

Published by the Faculty of Engineering, Khon Kaen University, Thailand

Attention-X: Enhancing the classification of natural attraction scenes with advanced attention mechanisms

Sujitranan Mungklachaiya and Anongporn Salaiwarakul*

Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok 65000, Thailand

Received 30 October 2024
Revised 13 February 2025
Accepted 18 March 2025

Abstract

This paper proposes the *Attention-X* method, which is an attention-based framework designed to address the challenges of interclass similarity and intraclass variance in natural scene classification tasks. The proposed method enhances pretrained convolutional neural networks (CNNs) by integrating an attention mechanism that selectively emphasizes salient and discriminative features, which enables the model to more effectively differentiate between visually similar scenes and manage variations within the same class. The proposed Attention-X method generates attention maps aligned with extracted features, integrating spatial representations with channel-wise relevance to overcome the limitations of the original deep features. This fusion enables the model to selectively amplify meaningful feature activations while suppressing irrelevant or redundant information. This improves the model's ability to distinguish between visually similar scenes and to handle variations within the same class. The proposed method was evaluated on the widely used SUN397, ADE20K, and Places365 benchmark datasets. The experimental results demonstrate that the proposed *Attention-X* method improves classification accuracy while maintaining competitive model complexity, outperforming several state-of-the-art methods. These findings highlight the effectiveness of the proposed method in real-world scenarios where subtle interclass differences and intraclass variability pose significant challenges.

Keywords: Scene classification, Deep learning, Convolutional neural networks, Attention mechanism

1. Introduction

In the tourism industry, scene classification is essential to automatically identify and categorize attraction scenes. It enhances tourist experiences by providing customized recommendations, comprehensive data, and appealing advertising, thereby leading to better decisions and greater satisfaction [1, 2]. In addition, the tourism operators can market lesser-known destinations to boost tourism economies and discover new attractions that are similar to popular sites, enabling personalized travel suggestions [3-5]. However, identifying natural attraction scenarios accurately is challenging because of the intraclass variance caused by weather, seasons, or viewpoints, which creates high variability within the same category [6, 7]. Interclass similarity further complicates the classification process because different categories frequently have similar visual features when images from distinct classes exhibit shared characteristics or overlapping components. The similarity of classes frequently results in classification errors because it is difficult for the model to differentiate between analogous properties. These constraints exacerbate the model's confusion and hinder its overall performance [8-10]; thus, these challenges must be addressed to enhance the reliability and effectiveness of natural attraction scene classification methods.

Deep learning (DL) techniques are highly effective in image classification tasks; however, when classifying images of highly complex natural scenes, using only DL models may be insufficient for accurate classification [11, 12]. Incorporating attention mechanisms can increase the model's ability to capture key points in images that are unique to each class. Attention mechanisms enable models to concentrate on crucial regions of an image, effectively ignoring irrelevant areas. This selective focus facilitates the extraction of more discriminative features that are essential for accurate classification. For example, models that integrate spatial and channel attention can highlight important features effectively while minimizing the impact of noise from less relevant regions [11, 13].

This paper discusses the classification of natural attraction scene images with a focus on strategies to overcome intraclass variance and interclass similarity. In addition, the Attention-X method is proposed to enhance classification performance by incorporating an attention mechanism into pretrained convolutional neural network (CNN) models (i.e., the Inception-V3, ResNet-50, and VGG-16 models). The proposed method is designed to enhance the model's efficacy by addressing the limitations of the deep features extracted solely from CNNs, which frequently struggle to focus on relevant information, thereby potentially increasing the accuracy and resilience in identifying natural attraction scenes. The goal of this study is to mitigate interclass similarity and intraclass variance, which are challenging issues in natural scene classification tasks due to scene similarity.

*Corresponding author.
Email address: anongporns@nu.ac.th

doi: 10.14456/easr.2025.30

The primary contributions of this study are summarized as follows.

1. We propose the Attention-X method to address challenges posed by interclass similarity and intraclass variance by emphasizing salient features using attention mechanisms, thereby making it versatile for various image classification tasks, particularly on datasets with subtle interclass similarity and intraclass variance.

2. We enhance pretrained CNN models by integrating attention mechanisms with different layer and filter configurations to generate an attention map that aligns with the extracted features from various models, thereby improving the scene classification performance.

3. The proposed Attention-X method integrates relevant channel-wise features with the spatial representation of the original image, which is extracted using a pretrained DL model. This fusion enables the Attention-X method to effectively focus on the most discriminative features based on the spatial structure of the input image. As a result, the model can classify images that exhibit intraclass variance and interclass similarity more effectively.

The proposed Attention-X method significantly improves classification accuracy, as demonstrated on a 16-class subset of the SUN397 dataset, highlight its ability to handle classification tasks that involve complex interclass and intraclass relationships, thereby paving the way for improved performance on challenging datasets and addressing real-world challenges.

2. Related work

Scene classification is particularly challenging due to interclass similarity, where scenes from different categories share common visual elements, and intraclass variance, where significant variations are present within the same category, which cause difficulties in terms of differentiating the meaning of the images. Initial attempts to improve the scene classification efficacy primarily depended on conventional feature extraction methods combined with machine learning approaches. Techniques include speeded up robust features extraction, K-means clustering, and linear discriminant analysis, which are frequently employed for scene image classification tasks. Note that these methods depend on handcrafted features. Thus, they work effectively in organized situations; however, they have significant drawbacks in complex and unstructured settings [14-16].The bag of visual words (BoVW) model was developed to address these issues. The BoVW model takes local features, e.g., those from the scale-invariant feature transform and histogram of oriented gradients, and groups them together into a “visual vocabulary.” However, the BoVW model neglects spatial information, which results in challenges in differentiating between images with analogous textures but varying configurations [16]. The spatial pyramid matching (SPM) approach was proposed to address these limitations. The SPM approach employs a hierarchical model to capture the spatial distributions of local information, which enhances the classification accuracy on various benchmark datasets, e.g., the MIT Indoor-67 and SUN397 datasets [14]. In addition, previous studies have demonstrated the performance of the BoVW model in land-use scenario classification by incorporating spatial interactions among local descriptors [17]. In addition to low-level feature descriptors, mid-level feature representations have emerged as a viable alternative to improve scene classification performance. These strategies aim to discover representative and discriminative sections or patches inside the images. A previous study [18] proposed an unsupervised technique to extract discriminative patches devoid of specified labels, yielding a more informative feature representation than conventional BoVW-based approaches. Hierarchical latent area models, as described in the literature [19], improved the ability to identify features in complex images by considering different spatial relationships at various scales.

Advancements in scene classification have progressed beyond feature extraction, including image segmentation and multilabel learning techniques to enhance accuracy and resilience. Image segmentation, especially with neural networks, has demonstrated significant efficacy in pixelwise semantic segmentation, enabling the distinction of essential scene elements, e.g., water, land, and sky [20]. In addition, multilabel learning techniques, e.g., the label correlation K-nearest neighbors method, have exhibited enhanced classification accuracy by examining interlabel correlations, thereby improving scene identification efficacy. Additionally, feature learning methodologies that use both labeled and unlabeled data have been devised to tackle the issue of redundant features in high-dimensional datasets, thereby diminishing the computing complexity and enhancing the overall classification accuracy [16, 21].

Traditional methods have provided valuable foundations; however, they frequently rely on handcrafted features and lack robustness against scene complexity. CNNs have revolutionized scene classification by automatically learning hierarchical representations from raw images [22, 23]. CNNs are widely used in scene representation and classification tasks, and CNN models, e.g., ResNet and VGG, have demonstrated promising results for both indoor and outdoor scenes. Utilizing pretrained CNN models significantly improves the speed and efficiency of recognizing large-scale natural scene images [24-26]. In addition, enhanced CNN techniques have been investigated to further improve the classification results. For example, a hierarchical Wasserstein CNN enhances scene classification performance by identifying interclass relationships and leveraging the hierarchical structure of images to minimize intraclass variance [27]. This method integrates advanced DL technologies to improve the classification accuracy. Despite the ability of CNN-based frameworks to perform feature extraction and image learning autonomously for individual classes, they frequently neglect shared objects or components among classes, resulting in a reduced ability to differentiate between classes with similar features [28].

To differentiate similar features, attention mechanisms are frequently incorporated into DL frameworks to extract the relevant features in an image [29-31]. Here, the goal is to extract distinctive and important features specific to each class while discarding nondiscriminative features before the classification process. This approach can reduce the interclass similarity problem because many images may contain similar components, which are common features found in multiple classes but are not distinctive features or objects for the classification of the specific class. Previous studies [32, 33] utilized attention mechanisms in conjunction with contextual data to refine the image representations used for classification and reduce the impact of intraclass variance and interclass similarity. To analyze the image features, a multiscale cross-attention mechanism combined with covariance pooling has been proposed [34]. Attention mechanisms can also extract salient objects, which are elements in images with distinct features that distinguish them from the background or other elements. Such mechanisms incorporate both spatial and channel attention to identify salient objects, thereby delineating discriminative image regions, particularly in scenarios where such regions are difficult to identify [29]. Another technique to improve the classification accuracy is a fusion strategy that combines diverse semantic and heterogeneous features using an adaptive weighted fusion algorithm. This method is based on extracting and learning features at different levels of detail from many sources in a hierarchical manner, which improves the ability to represent scenes and reduces the differences between visual and semantic elements. Here, the primary objective is to improve scene representation, harmonize visual-semantic disparities, integrate multigranularity features, and incorporate an attention mechanism to extract latent ontology features. In addition, a scene classification algorithm was proposed to minimize intraclass variation and maximize interclass distinctions, thereby offering improved solutions to the challenges associated with scene classification tasks [35]. Various attention mechanisms, e.g., the SE (Squeeze-and-Excitation Networks), CBAM

(Convolutional Block Attention Module), and ECA (Efficient Channel Attention) modules, have also been applied to refine fine-grained deep CNN (DCNN) image classification models, e.g., ResNet and VGGNet [36].

The literature review has demonstrated that although CNNs are proficient in classifying images, they encounter challenges when handling feature similarity issues. Even though attention mechanisms have simplified the process of grouping similar features, issues with intraclass variance and interclass similarity remain. The complexities of this problem are common in natural scenes. Thus, the goal of this study is to improve the image classification accuracy for images that face this challenge.

3. Proposed Attention-X method

This proposed Attention-X method is integrated into deep feature maps extracted from the CNN backbone network, emphasizing important image features to enhance the capacity of the pretrained CNN model to realize precise classification of natural scene images. The proposed Attention-X method, as shown in Figure 1, facilitates the extraction of relevant features required to distinguish each image while minimizing the irrelevant features required for categorization. This results in reduced class uniformity, which in turn leads to a decrease in interclass similarities. The selection of 16 natural attraction scene classes in this study was based on their representation, demonstrating significant variation within each class and similarities between the classes. The complexity of classifying specific categories arises from the variations within classes and the commonalities among classes, resulting in a significant reduction in the classification accuracy.

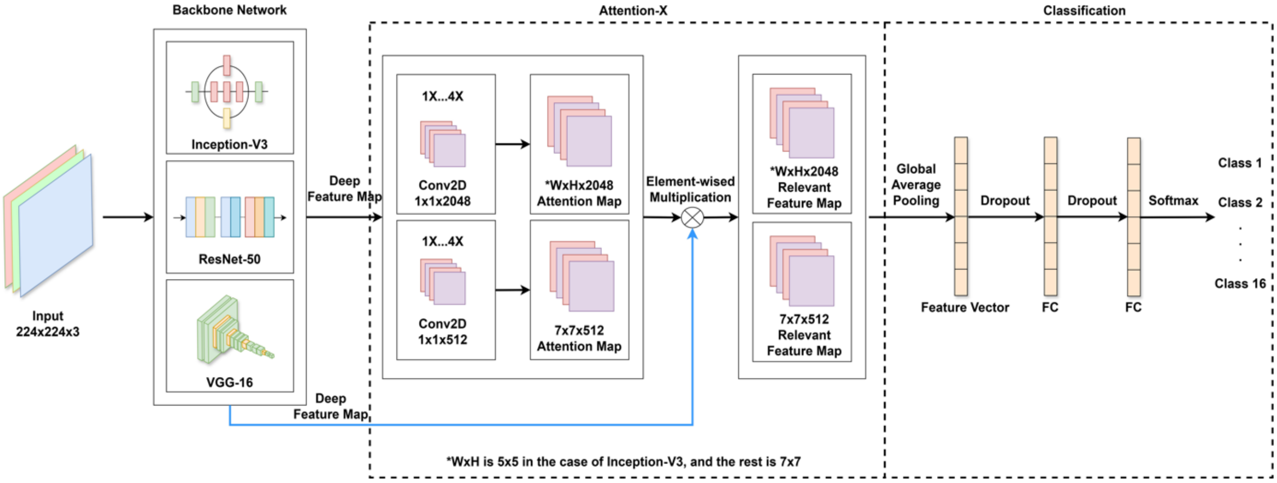


Figure 1 Architecture of proposed Attention-X method

3.1 Backbone network

Initially, we employed a pretrained CNN to extract high-level features from the input images (RGB format; 224 × 224 pixels). This method generated a deep feature map that captured the essential characteristics of each input image. Then, the Attention-X module employed the deep feature map that was obtained from the backbone pretrained CNN model. We performed an experimental comparative analysis of the backbone networks from three pretrained CNN models, i.e., the Inception-V3, ResNet-50, and VGG-16 models, which were selected for their widespread use in DL methods and their distinct architectures. The Inception-V3 model employs factorized convolutions and grid size reduction for computational efficiency, the ResNet-50 model utilizes residual connections to support deeper network training, and the VGG-16 model offers a simple, uniform architecture as a baseline. This diversity ensures a comprehensive evaluation of the proposed Attention-X method across different architectural paradigms, thereby allowing us to assess its effectiveness in enhancing performance across fundamentally different network designs. This also enables comparative analyses to identify the most effective model and configuration.

3.2 Attention-X module

The deep feature map extracted from the previous backbone network step was input to the Attention-X module. The purpose of this module is to improve the process of extracting features from the data and handling the data size. Here, the goal is to identify important regions in the images and provide a concise and informative representation of these features. The layer in attention mechanism is divided into X attention layers, referred to as Attention-X, to enhance the model's ability to focus on discriminative regions in the images. Each layer is constructed with an $Conv2D$ operation with a 1 × 1 filter size. Note that 1 × 1 convolutions do not capture spatial relationships across multiple pixels; however, they learn dependencies across different channels in the input feature map efficiently. Here, each 1 × 1 filter applies a linear transformation over the channels, which enables the model to refine and highlight important features.

In this study, the number of filters for the Inception-V3 and ResNet-50 models was set to 2048. The experiment also increased the original number of filters from 512 to 2048 for comparative analysis, and this model was designated VGG-16E. Note that the initial number of 512 filters in the VGG-16 model was utilized for the accuracy assessment. Although the 1 × 1 filters did not capture the spatial relationships across multiple pixels, they learned the feature relationships across different channels (depth) of the input feature map efficiently. Here, each 1 × 1 filter essentially provided a weighted summation over the channels, allowing the model to highlight particular features that were relevant to the attention mechanism. An attention map, denoted A, was generated by adding $Conv2D$ layers to the initial deep feature map F. This process is expressed in Eq. (1), where σ denotes the sigmoid function.

$$A = \sigma(ReLU((Conv2D_{1x1,N})^X(F))) \tag{1}$$

The convolutional layers, i.e., $Conv2D$, utilized a filter size of 1 × 1 and the $ReLU$ activation function, where N = 512 for the VGG-16 model and 2048 for the Inception-V3**,** ResNet-50, and VGG-16E models. The convolution operation was applied Xtimes, with $X \in \{1,2,3,4\}$ representing the number of $Conv2D$ layers, ranging from 1–4. The preceding layer generates a feature map that serves as the input to the subsequent layer, which comprises 2048 filters. The latter layer utilizes the output from the former layer as its input. The final result of processing the $Conv2D$ layers was the attention map A that integrated the extraction of the features, dimensionality reduction, and preservation of the spatial information in the images for utilization in the subsequent stage.

In the final stage, to create the relevant feature map, the attention maps were multiplied element-wise with the appropriate deep feature maps F recovered by the pretrained CNN models. This element-wise multiplication enables the model to emphasize the spatial and channel features simultaneously, as shown in Eq. (2), where $\otimes$ denotes element-wise multiplication.

$$R = F \otimes \mathrm{A} \tag{2}$$

This process combines diverse information from several feature maps that are considered important by the attention mechanism. The attention layer computes the mean value of the elements in each feature map channel, resulting in a compact feature vector that encapsulates the most significant information about the scene, which leads to the creation of a relevant feature map R with enriched feature representations. The relevant feature map effectively differentiates the discriminative features in the image using the attention mechanism, thereby resulting in a compact feature vector that encapsulates the most significant information about the scene.

In the proposed Attention-X method, 1 × 1 convolution layers are utilized to generate an attention map. This process integrates channel-wise relevant features with spatial information by applying the attention map, which encodes the weights representing the most significant channels at each pixel position, to the deep feature maps extracted by the pretrained CNN model. The pretrained CNN model employs convolutional operations to capture various hierarchical features, e.g., edges, textures, and high-level patterns, including shape, object parts, and spatial relationships, thereby preserving the spatial representation of the original image. As a result, the Attention-X model can effectively focus on the most discriminative features according to the spatial structure of the input image, resulting in the extraction of more representative features that are suitable for classifying similar scene images.

3.3 Classification module

The model's classification module was designed to perform dimensional reduction and regularization, in addition to the classification process. The classification module obtains relevant feature maps from the Attention-X module, and then global average pooling diminishes the spatial dimensions and transforms the feature maps into a flattened 2048-dimensional feature vector. However, for VGG-16, the feature vector has 512 dimensions. Here, a dropout layer with a rate of 0.5 was implemented to mitigate overfitting.

The feature vector was then converted into a 2048-dimensional fully connected (FC) layer to capture the intricate interactions among the features with $ReLU$ activation. A supplementary dropout layer was incorporated before the output FC layer, which included 16 units matching the number of classes. The output FC layer in the classification process utilized the SoftMax activation function to determine the probability of each potential scene categorization.

4. Experimental settings

The experimental setting comprised two sections. The first section focused on the dataset, and the second section examined the proposed Attention-X method, which integrates the attention mechanism into DL to assess the performance of the Inception-V3, ResNet-50, and VGG-16 models via comparative analysis. Here, attention layers were employed to improve the extraction of relevant features from the three pretrained Inception-V3, ResNet-50, and VGG-16 CNN models. In this experiment, each pretrained CNN model constituted the foundation for the primary feature extraction procedure.

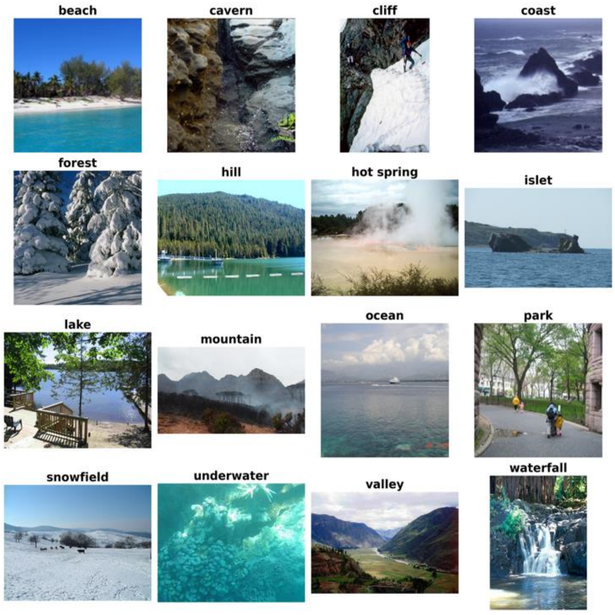


Figure 2 Examples of 16 different natural scene classes from the SUN397 dataset

4.1 Dataset

To evaluate the efficacy of integrating attention mechanisms with deep features, we selected a subset of natural attraction scene images from the SUN397 [37] benchmark dataset, spanning 16 distinct categories, e.g., beach, cliff, hill, and hot spring. These image classes were utilized in experimental investigations performed to overcome the challenges in image recognition and classification tasks, including intraclass variability, interclass similarities, and images lacking distinctive objects. Examples of images from each class in the dataset are shown in Figure 2, and the class distribution is shown in Figure 3.

In this experiment, the dataset was split into training and test sets with 4,801 and 1,202 images, respectively, following an 80:20 ratio for each class, as shown in Table 1. This distribution ensured that each class has a proportional representation in both the training and test sets.

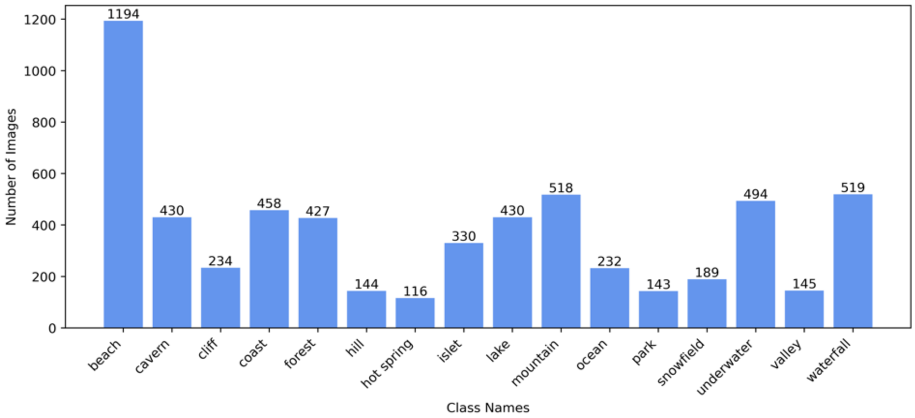


Figure 3 Distribution of 16 classes in the SUN397 dataset

Table 1 Experimental dataset details

Class name	Number of images	Training set	Test set
beach	1,194	955	239
cavern	430	344	86
cliff	234	187	47
coast	458	366	92
forest	427	342	85
hill	144	115	29
hot spring	116	93	23
islet	330	264	66
lake	430	344	86
mountain	518	414	104
ocean	232	186	46
park	143	114	29
snowfield	189	151	38
underwater	494	395	99
valley	145	116	29
waterfall	519	415	104

4.2 Model settings

An experimental model configuration was designed to train and validate the proposed method and assess its effectiveness. In addition, a comparative analysis was performed to examine the effect of integrating attention mechanisms into the baseline backbone pretrained CNN models (Inception-V3, ResNet-50, and VGG-16) and identify the most optimal model. The additional layers were incorporated into the attention mechanism, denoted Attention-X, where X is the number of layers appended to the attention mechanism. Here, the objective was to verify the most effective layer configuration for the attention mechanism, spanning 1 to X, with varying numbers of filters in each layer. The number of filters in the attention layers was set to match the number of filters in the final feature map of each baseline model, which allowed the integration of the attention mechanisms without modifying the extracted features. Note that the Inception-V3 and ResNet-50 models have 2048 filters, and the VGG-16 model has 512 filters; thus, two configurations were investigated for the VGG-16 model, i.e., one with 512 filters in the attention layers (VGG-16) and one with 2048 filters (VGG-16E).

In this study, 1–4 attention layers (X) were applied to each backbone pretrained CNN model. Thus, two Attention-X configurations were created. In the first configuration, the filter count in each attention layer corresponded with the number of filters in the final feature map of the backbone model, labeled 1X–4X. Here, the Inception-V3 and ResNet-50 models contained 2048 filters, and the VGG-16 contained 512 filters. In the second configuration, the model was specific to VGG-16E, where the initial number of 512 filters in each attention layer was increased in the standard VGG-16 architecture to 2048 filters. In the configuration with three attention layers, the model complexity was reduced by halving the number of filters in the second layer. This reduction is denoted by appending “r” to the configuration names, e.g., 3Xr. Details of each experimental configuration are shown in Table 2.

The motivation for varying the number of attention layers was to create attention maps that better emphasize important features and capture more essential relationships for classification.

Table 2 Experimental settings

Backbone pretrained CNN model : Inception-V3, ResNet-50, VGG-16E					
Number of attention layers (X)	1X	2X	3X	3Xr	4X
Number of filters in each layer	2048	2048 2048	2048 2048 2048	2048 1024 2048	2048 2048 2048 2048
Backbone pretrained CNN model : VGG16					
Number of attention layers (X)	1X	2X	3X	3Xr	4X
Number of filters in each layer	512	512 512	512 512 512	512 256 512	512 512 512 512
Loss function (cross-entropy), optimizer (Adam), batch size (32), learning rate (0–1.00e−04), dropout rate (0.5), dense layer size (2048), strides (1,1), padding (zero padding)					

The initial layer identified the basic patterns, and the subsequent layers refined the attention map to produce a more detailed representation. In addition, the number of filters was varied in the experiment to evaluate whether models with fewer filters could still realize optimal performance. For example, a 2048-1024-2048 configuration (3Xr for Inception-V3, ResNet-50, and VGG-16E) was tested. VGG-16 was also tested with a 512-256-512 configuration to evaluate the effects of the reduced filter dimensions on the performance of the model.

In Attention-X, the resulting attention map from the X layers was multiplied element-wise with the deep feature map from the backbone network of each pretrained CNN model. Global average pooling was employed to reduce the dimensionality of the data and provide an appropriate feature vector for the dense layers. This helped mitigate the overfitting of the model. To enhance the resistance to overfitting, we included dropout layers following both the dense layers and global average pooling. Here, the dropout layers employed a random process to deactivate 50% of the nodes prior to forwarding the inputs to the classification layer.

The effectiveness of the models was assessed using several evaluation metrics, including accuracy, precision, recall, and the F-measure. In addition, the model size and the number of Giga Floating Point Operations (GFLOPs) were analyzed to evaluate the computational complexity of each model.

Model accuracy, calculated by Eq. (3), is the evaluation metric that measures the proportion of correctly classified instances among the total number of instances, and reflects the overall effectiveness of the model in both identifying positive and negative cases correctly.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Precision, as defined in Eq. (4), quantifies the ratio of correctly predicted positive instances to the total number of predicted positive instances:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Recall evaluates the model’s ability to correctly identify actual positive instances, using Eq. (5).

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

where:

TP (True Positives) refers to correctly predicted positive samples,
FP (False Positives) refers to negative samples incorrectly predicted as positive,
FN (False Negatives) refers to positive samples incorrectly predicted as negative.

The F-measure, presented in Eq. (6), is the harmonic mean of precision and recall, providing a balanced metric of both:

$$F - measure = 2x\frac{Precision\ x\ Recall}{Precision + Recall} \tag{6}$$

Furthermore, the experiment assessed the model size and computational complexity, expressed in GFLOPs (Giga Floating Point Operations), which indicates the total number of floating-point operations required to process a single input. The FLOPs were calculated using Eq. (7):

$$FLOPs = profiler(convert\_variables\_to\_constants(tf.function(M_{Keras}, S_{input}))) \tag{7}$$

where

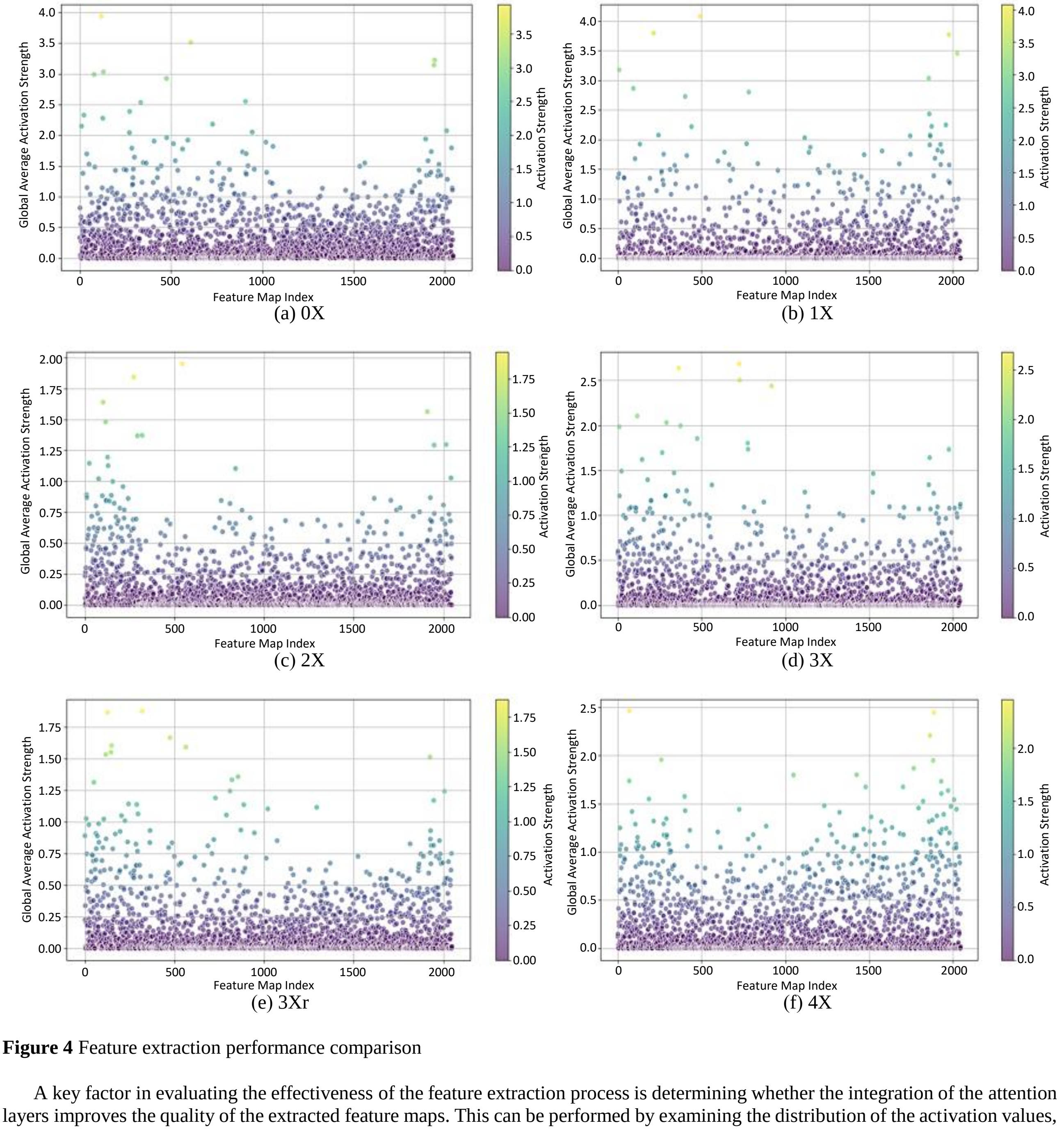
M_{Keras} is the original model defined in $Keras$,
$tf.function(M_{Keras}, S_{input})$ represents the transformation of the model into a TensorFlow static computation graph with the specified input shape,
$profiler$ is used to analyze the frozen computation graph and report the total number of operations performed in each layer.

5. Experiment results and discussion

5.1 Attention-X feature representation: layer-wise comparison

Based on the experimental settings described in Table 2, we performed experiments to integrate the deep feature maps extracted using a CNN pretrained model with attention maps generated using varying numbers of attention layers (X). Here, the objective was to determine the optimal number of attention layers that yielded the most descriptive and representative features. The impact of integrating different attention layer configurations is demonstrated through the combination of the Inception-V3 model with these variations, as shown in the scatter plots in Figure 4.

The scatter plots illustrate the global average pooling activation strength, comparing the feature extraction performance across different numbers of attention layers (X), i.e., 0X, 1X, 2X, 3X, 3Xr, and 4X. Among these configurations, the 0X configuration represents the deep feature maps extracted from the pretrained model without additional attention layers.



(a) 0X (b) 1X

(c) 2X (d) 3X

(e) 3Xr (f) 4X

Figure 4 Feature extraction performance comparison

A key factor in evaluating the effectiveness of the feature extraction process is determining whether the integration of the attention layers improves the quality of the extracted feature maps. This can be performed by examining the distribution of the activation values, ensuring a well-balanced spread that avoids excessive clustering at extreme low or high values. In addition, an unusually high maximum activation strength may indicate overfitting, where the model overemphasizes specific features while disregarding other features. The scatter plots in Figure 4 show the correlation between the feature index (x-axis; ranging from 0–2047) and the average activation strength (y-axis) derived by global average pooling. Here, higher activation values indicate that the associated feature maps are more influential in the model's decision-making process. The data points are color-coded, ranging from purple to yellow, indicating varying levels of activation strength. The dark purple points represent low activation values (close to zero), signifying characteristics that are infrequently employed by the model, and light green to yellow points indicate elevated activation strength, implying that these features influence the model’s predictions substantially. An investigation of the feature extraction from sample images indicated that the models with four attention layers (4X), as shown in Figure 4(f), provided the most effective feature representation. The scatter plot for the 4X configuration exhibits a balanced distribution of the activation strength, with both low and high values evenly spread across the feature maps. This signifies the model’s ability to extract diverse features, thereby enhancing its effectiveness in representing images for classification. As shown in Figure 4(e), the 3Xr configuration, which incorporates three attention layers while reducing the filter count in the intermediate layer, ranks second in performance. Its distribution pattern resembles that of 4X; however, it exhibits a slightly lower feature diversity. In contrast, Figure 4(a) shows the 0X configuration, which lacks attention layers. As can be seen, it exhibits a balanced distribution of the activation intensity; however, its performance remains modest. The absence of prominent and defining characteristics limits its effectiveness in feature representation. In contrast, the 1X configuration, as shown in Figure 4(b), exhibits clear signs of overfitting. Some feature maps have extremely high activation values (approaching 4.0), which suggests that the model excessively focused on a limited set of features while neglecting other features. This imbalance results in ineffective utilization of features.

In addition, the 2X and 3X configurations demonstrated the least efficient feature extraction capabilities, with the 3X configuration exhibiting the most significant deficiencies. The corresponding scatter plot indicates that most of the activation strength values were persistently low, lacking any distinctly prominent features. This indicates that the feature extraction technique in certain settings inadequately captures essential information, leading to unsatisfactory feature representation. These findings highlight the significance of an ideal number of attention layers in terms of maintaining feature variety, mitigating overfitting, and improving classification performance.

5.2 Experimental results

The experiment investigated various configurations for the baseline Inception-V3, ResNet-50, and VGG-16 model (Table 2) to determine the optimal configuration for each model. This study investigated the effect of increasing the number of attention layers to identify the optimal layer count that achieved the best accuracy. Initially, the experiment examined whether increasing the number of layers improved the accuracy. For the ResNet-50 and VGG-16 models, the accuracy improved as the number of layers increased from 2X to 3X; however, the opposite effect was observed for the Inception-V3 and VGG-16E models. Motivated by this finding, the number of layers was further increased to 4X to evaluate the optimal number of layers for each model and determine whether additional layers would yield further improvements. However, the accuracy of some models was reduced when the layers were increased from 3X to 4X, which indicates that the model complexity at 4X may be excessive. Then, we hypothesized that reducing the complexity of the 3X configuration may yield a simpler model with comparable performance. Thus, a reduced version, i.e., the 3Xr configuration, was tested, and the results, as shown in Figure 5, demonstrated the effectiveness of this approach.

In addition, the accuracy curves shown in Figure 5 indicate how the training and validation accuracy evolved as the epochs proceeded. The results suggest that certain models, e.g., the Inception-V3 (3Xr and 4X), ResNet-50 (3X), and VGG-16 (4X) models, achieved better generalizability, as evidenced by the smaller gap between the training and validation accuracy. This highlights the importance of parameter tuning, where adjusting the number of attention layers influences the performance of the models. These findings reinforce that increasing the number of layers does not always improve accuracy. Instead, a well-balanced architecture is required to prevent overfitting and unnecessary complexity.

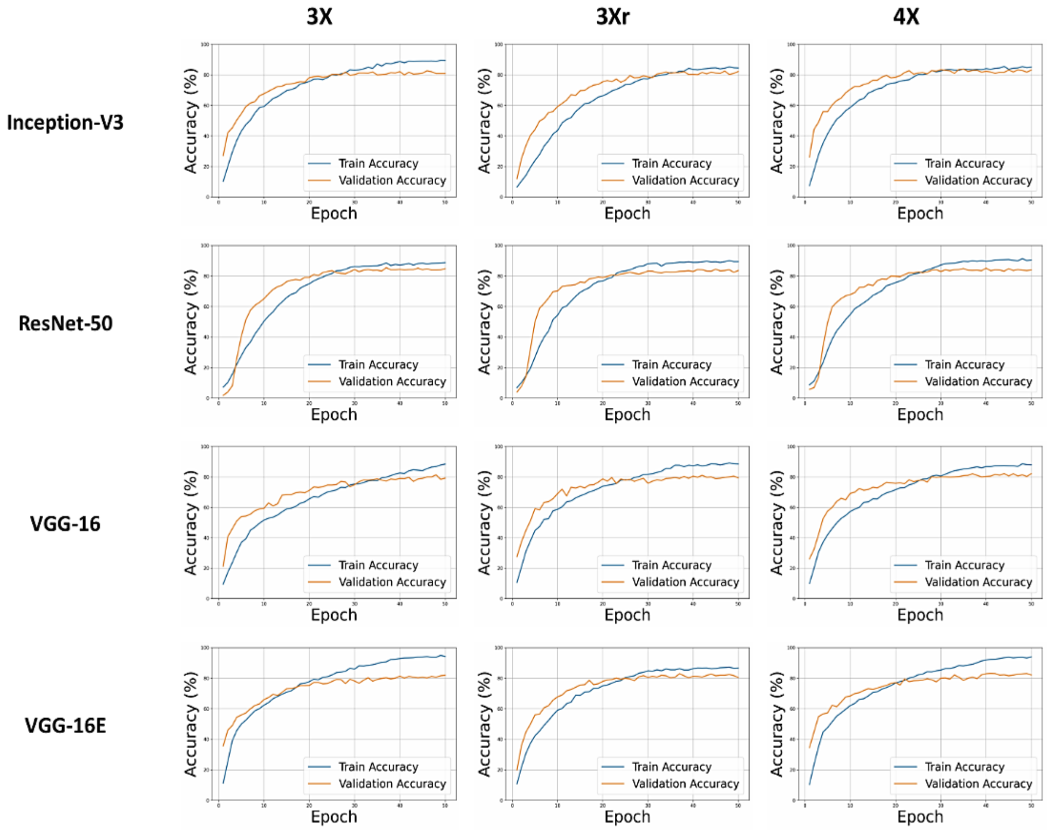


Figure 5 Training and validation accuracy across different attention layer configurations

Table 3 shows a comprehensive evaluation of each optimized model, highlighting the impact of the attention layers on the classification performance. The baseline CNN models encountered challenges in scene differentiation due to the intraclass variations and interclass similarities because they classified images without considering the shared elements among the classes, thereby making it difficult to extract distinctive features. For example, beaches, forests, and gardens may all contain trees, leading to misclassification if trees are considered as defining features. Incorporating attention layers improves the feature discrimination, particularly in the Inception-V3 model, where the performance remained suboptimal until the 3Xr configuration, which enhanced accuracy by reducing the number of filters in the intermediate layer. The highest accuracy of 84.61% was achieved with the 4X configuration, surpassing its baseline of 82.86%. Similarly, the ResNet-50 model demonstrated a consistent upward trend from a baseline of 80.45%, peaking at 84.03% with the 3X configuration. Although the 3Xr configuration caused a slight decline in accuracy, the performance recovers at 4X attention. The VGG-16 model exhibited performance fluctuations, demonstrating degradation with the 2X and 3Xr configurations, but improving significantly with the 4X configuration, reaching 82.28%, indicating that the additional attention layers benefited this model. In contrast, the VGG-16E model with 2048 filters demonstrated greater volatility than VGG-16, exhibiting substantial drops with the 1X and 3X configurations, but recovering with the 3Xr and 4X configurations, which suggests that reducing the number of attention layers introduces instability while higher levels enhance performance.

Table 3 Comparison of model accuracy and complexity

Model	**Number of**	**Learning**	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**Model size**	**GFLOPs**
	attention layers (X)	**rate**	**(%)**	**(%)**	**(%)**	**(%)**	**(MB)**	
Inception-V3	0	4.00e−06	82.86	83.22	82.86	82.71	99.30	5.70
	1	1.00e−04	82.36	83.35	82.36	82.39	115.31	5.91
	2	1.00e−04	82.03	82.48	82.03	81.76	131.32	6.12
	3	1.00e−04	80.78	81.77	80.78	80.42	147.33	6.33
	3r	2.00e−05	84.03	84.76	84.03	84.16	131.32	6.12
	4	**4.00e−06**	**84.61**	**84.61**	**84.53**	**84.42**	**163.34**	**6.54**
ResNet-50	0	1.00e−04	80.45	82.36	80.45	80.65	106.11	7.74
	1	4.00e−06	83.86	84.34	83.86	83.88	122.12	8.15
	2	2.00e−05	83.19	84.36	83.19	83.39	138.13	8.56
	3	6.40e−09	84.03	84.14	84.03	83.91	154.14	8.97
	3r	2.00e−05	82.70	83.68	82.70	82.70	138.13	8.56
	4	2.00e−05	83.61	83.97	83.61	83.45	170.14	9.40
VGG-16	0	2.00e−05	81.20	81.00	81.20	80.84	60.26	30.72
	1	2.00e−05	82.20	82.82	82.20	82.01	61.27	30.74
	2	1.00e−04	78.70	80.94	78.70	78.37	62.27	30.77
	3	2.00e−05	81.45	80.94	81.45	81.03	63.27	30.79
	3r	1.00e−04	76.96	77.74	76.96	76.46	62.27	30.77
	4	2.00e−05	82.28	82.14	82.28	81.99	64.27	30.82
VGG-16E	1	1.00e−04	77.29	78.39	77.29	77.13	80.28	30.93
	2	2.00e−05	82.20	82.56	82.20	82.15	96.29	31.34
	3	1.00e−04	77.37	77.09	77.37	76.53	112.30	31.75
	3r	2.00e−05	82.03	82.04	82.03	81.76	96.29	31.34
	4	2.00e−05	81.45	81.37	81.45	81.11	128.30	32.16

A comparative examination of the results presented in Table 3 shows that the attention layers generally enhanced the model performance across diverse architectures, with notable improvements observed with the ResNet-50 and Inception-V3 models. Incorporating additional attention layers into the ResNet-50 model significantly enhanced the performance, exhibiting a distinct increasing trajectory. Excessive attention layers or inadequate tuning, e.g., the VGG-16 model at lower levels, may result in reduced accuracy. With the 3Xr configuration, reducing the number of filters in the second attention layer appeared to reduce the prediction accuracy. This finding indicates that reducing the number of filters impacted the model's ability to learn image features effectively. In addition, using an equal number of filters across all layers facilitated better learning, as demonstrated in Table 3 for the 3X and 4X configurations, except in the case of the Inception-V3 model. However, the VGG-16 model was more susceptible to variations in the number of layers and the filter sizes. The ResNet model's skip connections were more adept at managing increased complexity; however, the VGG model's deep sequential layers may falter under excessive attention processes.

In terms of computational cost and model complexity, the complexity of each model changed considerably with the added attention layers, thereby making it crucial to balance performance and computational costs. The Inception-V3 model demonstrated that adding more attention layers increased both the model size and the computational cost; however, the performance improvement was not strictly linear. The 3Xr configuration was particularly noteworthy for achieving a balance between higher accuracy and reduced complexity, making it an efficient choice. With the 4X configuration, although the accuracy increased slightly, the added complexity was considerable.

The Inception-V3 and ResNet-50 models demonstrated the most consistent improvements with the additional attention layers; however, the complexity of these models increased proportionally. In contrast, the VGG-16 and VGG-16E models exhibited more instability in performance with increased attention layers, and their GFLOPs remained high due to the inherent depth of the architecture. For all models, the 3Xr configuration generally offered a good balance of performance and efficiency.

The experimental results demonstrate that attention layers typically improve performance although the impact differs by architecture. For example, the ResNet-50 model exhibited enhanced stability when additional attention layers were incorporated.

5.3 Addressing intraclass variance and interclass similarity with Attention-X

In the classification of natural scene images, intraclass variance is a common challenge because variations in different factors, e.g., orientation, seasons, and color tone, can substantially impact model performance. Interclass similarity is also a challenge because numerous natural scene classes share similar elements or objects. In addition, landscape images require meticulous attention to both the foreground and distant background details because these factors influence the classification accuracy.

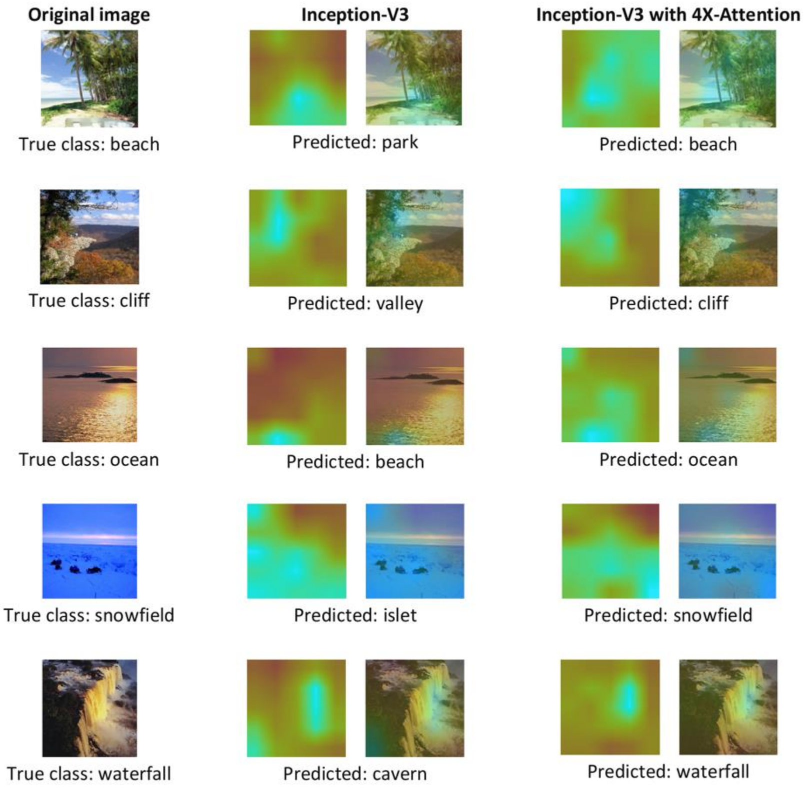


Figure 6 Feature map heatmap and highlighted regions of challenging images

Figure 6 illustrates these challenges by highlighting the intraclass diversity within the beach class, which lacks distinctive sand elements. This diverges from typical beach images, with trees emerging as the predominant feature. This results in a certain level of interclass similarity with images from the park class. Similarly, the brownish hue of an ocean image, which differs from other ocean class images, may result in erroneous classification as a beach. When analyzing the feature map heatmaps and highlighted regions of the challenging images extracted using the Inception-V3 and Inception-V3 models with 4X-Attention configuration, as well as the resulting predictions, the highlighted regions reveal that the lightest colors represent the features the model considers the most important, and the darker areas are considered less relevant for the prediction tasks. This indicates that the proposed Attention-X method helps the models identify salient regions within the images, particularly for the Inception-V3 and Inception-V3 models with 4X-Attention, thereby leading to ideal results. Thus, the model handles intraclass variation and interclass similarity more adeptly, surpassing the efficiency of conventional CNN models.

5.4 State-of-the-art comparison

This section presents a comprehensive comparison of our best model, based on the proposed Attention-X method, Inception-V3-4X, which was trained and fine-tuned on 16-class natural scene images from the SUN397 dataset, against other state-of-the-art models. This comparison includes different methodological approaches, i.e., CNN-based architectures and CNNs with attention mechanisms, including AlexNet [38], Places-CNN [38], and SAS [39]. The SAS model, which integrates a CNN with semantic segmentation to leverage object relationships in the image for classification, is further divided into three variations, i.e., models pretrained on the SUN397 and Places365 [38] datasets, which were directly evaluated on the test set of the evaluation dataset without additional training or fine-tuning processes. In addition, we evaluated SAS* and AlexNet*, which were trained and fine-tuned on 16 resembling classes from the SUN397 dataset, the same dataset used in our experiments, to evaluate their performance when trained on a targeted set of classes. In addition, we tested the Places-CNN model pretrained on the Places365 dataset without additional training or fine-tuning processes.

To evaluate the performance of CNNs with attention mechanisms, the Inception-V3-4X model was compared with the CBAM [40] and SENet [41] models, both of which integrate attention mechanisms in conjunction with CNN-based feature extraction, comparable to the proposed Attention-X method. The CBAM model integrates both channel and spatial attention; however, the SENet model relies solely on channel attention, aligning with the design of the proposed model. The architectures of both models were presented using ResNet-50 as the backbone network. To facilitate a fair evaluation, the CBAM and SENet models were trained and fine-tuned according to their original architectures using the 16-class subset of the SUN397 dataset.

This evaluation was performed using several benchmark datasets, i.e., the SUN397, ADE20K [42], and Places365 datasets. Note that Places365 is a large-scale dataset; thus, we selected 100 images per-class for testing to ensure that the number of test samples was generally aligned with the average test data size of the SUN397 dataset, which served as the training dataset for our model. Emphasizing both accuracy and model size, Table 4 shows the strengths and limitations of each model based on accuracy and model size.

The comparative analysis of the Inception-V3-4X* model (referring to the Inception-V3 model with 4X-Attention) against the CNN architecture methods highlighted their strengths and weaknesses in terms of accuracy and computational efficiency. The

Inception-V3-4X* model, fine-tuned on 16 resembling classes from the SUN397 dataset, consistently outperformed the other models, obtaining the highest accuracy on both the SUN397 (84.61%) and ADE20K (86.27%) datasets; however, this model exhibited only moderate performance on the Places365 dataset, in which the dataset was tested 100 images per class. This implies that its architecture and fine-tuning process are slightly limited in capturing the relevant features required for scene classification across diverse datasets.

Table 4 Comparison of state-of-the-art models on SUN397, ADE20K, and Places365 test datasets

Method	Model	Trained dataset	Backbone	Size (MB)	SUN397 (%)	ADE20K (%)	Places365 (%)
CNN-based architectures	SAS [39]	SUN397	ResNet-18	107.76	77.37	72.19	54.50
	SAS [39]	Places365	ResNet-18	107.50	57.96	61.66	59.50
	SAS* [39]	SUN397	ResNet-18	139.23	80.95	85.20	50.71
	AlexNet* [38]	SUN397	AlexNet	217.71	78.20	84.31	45.36
	Places-CNN [38]	Places365	ResNet50	92.53	70.38	67.19	69.79
CNNs with attention mechanisms	CBAM* [40]	SUN397	ResNet50	90.00	85.11	88.77	51.64
	SENet* [41]	SUN397	ResNet50	92.01	64.73	57.75	34.43
	Inception-V3-4X*	**SUN397**	**Inception-V3**	**163.34**	**84.61**	**86.27**	**51.07**

* Trained and fine-tuned 16 similar classes from the SUN397 dataset using the same data as the Attention-X experiments to ensure a fair comparison of the feature extraction performance with our model.

Similarly, the model SAS*, optimized for 16 classes from SUN397, demonstrated robust performance on the SUN397 (80.95%) and ADE20K (85.20%) datasets; however, the model SAS* underperformed on the Places365 dataset, suggesting restricted generalizability to datasets beyond its training scope. Despite optimization over 16 identical classes, the Inception-V3-4X model surpassed AlexNet* on most datasets, perhaps because of its antiquated design and diminished ability to capture complex scene features.

Places-CNN with a ResNet-50 backbone obtained the highest accuracy on the Places365 dataset in the 100 images/class setting (69.79%), reflecting its alignment with the dataset it was pretrained on. However, its lower performance on the SUN397 and ADE20K dataset highlights the trade-off between specialization and generalization across datasets.

From a computational efficiency perspective, the model Inception-V3-4X* struck an effective balance between model size and accuracy, with a moderate size (163.34 MB) and consistently high performance across the evaluated datasets. In contrast, the AlexNet* model, despite its larger size (217.71 MB), did not significantly outperform smaller models like SAS*, highlighting the advantages of the Inception-V3 model's modern architecture in terms of balancing computational complexity and performance.

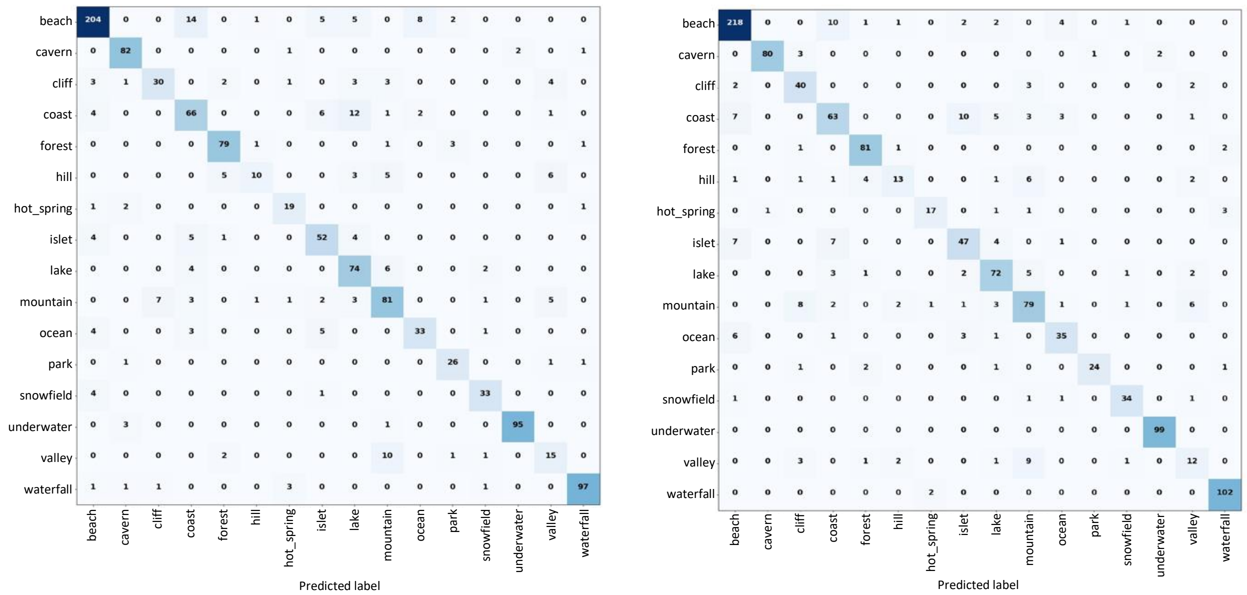
The comparison of the Inception-V3-4X* and CNNs with attention mechanism models (i.e., CBAM and SENet) highlights key differences in the accuracy and efficiency of the scene classification task. While the CBAM and SENet models leverage attention mechanisms, the Inception-V3-4X* model achieved comparable or superior accuracy on the SUN397 (84.61%) and ADE20K (86.27%) datasets, outperforming the SENet model and closely matching the CBAM model despite lacking explicit attention modules. On the Places365 dataset, the CBAM model performed the best (51.64%), while the Inception-V3-4X* model remained competitive (51.07%), demonstrating its ability to classify structured scenes effectively.

Computationally, the Inception-V3-4X* model is larger (163.34 MB) than the CBAM (90.00 MB) and SENet (92.01 MB) models; however, it maintained high accuracy, suggesting efficient feature extraction without attention mechanisms. However, the CBAM model's spatial and channel attention enhances its generalizability, which indicates that adding attention layers or hybridizing with transformers could further improve the Inception-V3-4X* model.

Overall, the Inception-V3-4X* model demonstrated a balanced trade-off between accuracy and computational efficiency. Although its accuracy was lower than that of the CBAM model across all datasets, it outperformed the SENet model, which relies solely on channel attention. This finding suggests that the Inception-V3-4X* model effectively captures critical scene features through its channel attention mechanism, even without explicitly incorporating spatial attention, as seen in the CBAM model. However, the CBAM model benefits from integrating both spatial and channel attention, which enhances its ability to distinguish complex scenes and contributes to its superior performance across all datasets. Nonetheless, the Inception-V3-4X* model obtained results that were comparable to those of the CBAM model and surpassed the SENet model while utilizing only channel attention. This highlights the effectiveness of the proposed Attention-X method, which seamlessly integrates with the multiscale feature extraction architecture and deep convolutional layers of the Inception-V3 model, thereby enabling robust feature learning without strong dependence on spatial attention. From the model size perspective, the Inception-V3-4X* model is considerably larger (163.34 MB) than the CBAM (90.00 MB) and SENet (92.01 MB) models. Generally, a larger model size reflects an increased capacity for learning complex features; however, the CBAM model achieved higher accuracy with a more compact structure, which indicates that it may be more architecturally efficient in terms of the performance-to-size ratio.

5.5 Discussion

The Attention-X model, which implements an attention mechanism to focus on important salient features while disregarding irrelevant features, effectively addressed the intraclass variance and interclass similarity issues, resulting in more accurate classification results. The confusion matrix shown in Figure 7, which compares the conventional Inception-V3 model with the proposed Inception-V3-4X model (Inception-V3 with four attention layers) that utilized all 2048 filters, reveals that the Inception-V3-4X model significantly improved the classification accuracy in scenes with expansive backgrounds and predominantly uniform color schemes, e.g., beaches, cliffs, forests, hills, underwater, and waterfall scenes. Upon closer examination of the findings for different classes e.g., caverns, hot springs, islets, lakes, mountains, parks, and valleys, we found that, despite its overall superior accuracy, the model exhibited lower accuracy compared with the model on a per-class basis. Misclassifications were greater in natural scenes, which have prominent backgrounds and an inadequate number of distinguishing foreground objects, compared with scenarios without attention layers.



	beach	cavern	cliff	coast	forest	hill	hot_spring	islet	lake	mountain	ocean	park	snowfield	underwater	valley	waterfall
beach	204	0	0	14	0	1	0	5	5	0	8	2	0	0	0	0
cavern	0	82	0	0	0	0	1	0	0	0	0	0	0	2	0	1
cliff	3	1	30	0	2	0	1	0	3	3	0	0	0	0	4	0
coast	4	0	0	66	0	0	0	6	12	1	2	0	0	0	1	0
forest	0	0	0	0	79	1	0	0	0	1	0	3	0	0	0	1
hill	0	0	0	0	5	10	0	0	3	5	0	0	0	0	6	0
hot_spring	1	2	0	0	0	0	19	0	0	0	0	0	0	0	0	1
islet	4	0	0	5	1	0	0	52	4	0	0	0	0	0	0	0
lake	0	0	0	4	0	0	0	0	74	6	0	0	2	0	0	0
mountain	0	0	7	3	0	1	1	2	3	81	0	0	1	0	5	0
ocean	4	0	0	3	0	0	0	5	0	0	33	0	1	0	0	0
park	0	1	0	0	0	0	0	0	0	0	0	26	0	0	1	1
snowfield	4	0	0	0	0	0	0	1	0	0	0	0	33	0	0	0
underwater	0	3	0	0	0	0	0	0	0	1	0	0	0	95	0	0
valley	0	0	0	0	2	0	0	0	0	10	0	1	1	0	15	0
waterfall	1	1	1	0	0	0	3	0	0	0	0	0	1	0	0	97

Predicted label

(a) Inception-V3

	beach	cavern	cliff	coast	forest	hill	hot_spring	islet	lake	mountain	ocean	park	snowfield	underwater	valley	waterfall
beach	218	0	0	10	1	1	0	2	2	0	4	0	1	0	0	0
cavern	0	80	3	0	0	0	0	0	0	0	0	1	0	2	0	0
cliff	2	0	40	0	0	0	0	0	0	3	0	0	0	0	2	0
coast	7	0	0	63	0	0	0	10	5	3	3	0	0	0	1	0
forest	0	0	1	0	81	1	0	0	0	0	0	0	0	0	0	2
hill	1	0	1	1	4	13	0	0	1	6	0	0	0	0	2	0
hot_spring	0	1	0	0	0	0	17	0	1	1	0	0	0	0	0	3
islet	7	0	0	7	0	0	0	47	4	0	1	0	0	0	0	0
lake	0	0	0	3	1	0	0	2	72	5	0	0	1	0	2	0
mountain	0	0	8	2	0	2	1	1	3	79	1	0	1	0	6	0
ocean	6	0	0	1	0	0	0	3	1	0	35	0	0	0	0	0
park	0	0	1	0	2	0	0	0	1	0	0	24	0	0	0	1
snowfield	1	0	0	0	0	0	0	0	0	1	1	0	34	0	1	0
underwater	0	0	0	0	0	0	0	0	0	0	0	0	0	99	0	0
valley	0	0	3	0	1	2	0	0	1	9	0	0	1	0	12	0
waterfall	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	102

Predicted label

(b) Inception-V3-4X

Figure 7 Classification confusion matrixes of the Inception-V3-4X and baseline models

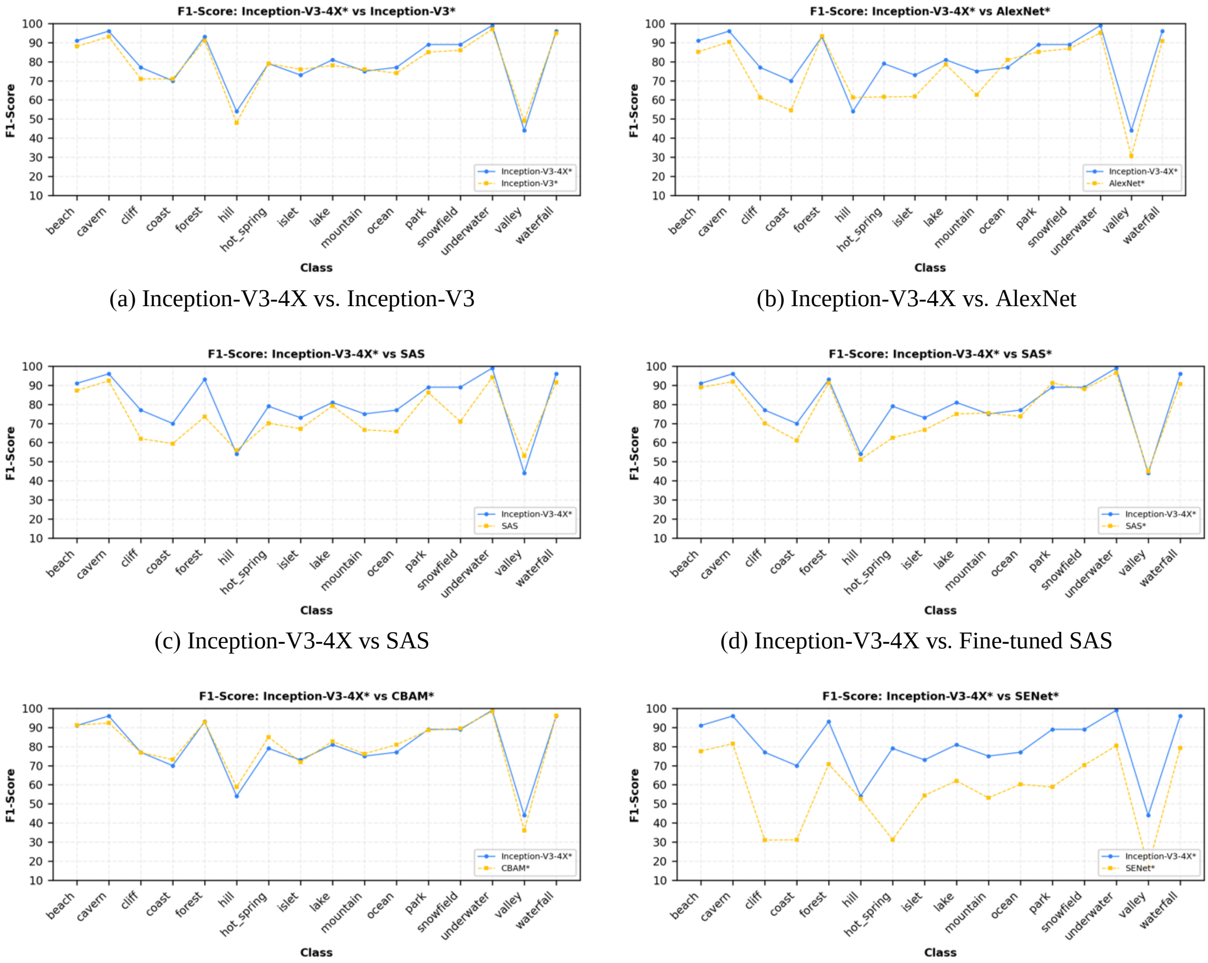
This suggests that although the model demonstrated improved proficiency in classifying scenes within specific classes it struggled with similarities between other categories. For example, distinguishing between scenes from different classes that share similar colors, components, and objects, e.g., the coast and ocean, valley and mountain, and cliffs with waterfalls where streams flow down from high cliffs, can lead misclassification. Here, the attention mechanism discriminately assigns higher priority to significant elements of the image while disregarding insignificant aspects, which implies that the model ignores less important elements of the image, and this can impact its understanding of the overall context. This is especially true when dealing with objects or areas that are challenging to identify, as demonstrated in our experiments with the Inception-V3-4X model, which employed four layers with filters of dimensions 2048 × 2048 × 2048 × 2048. Maintaining 2048 filters across all layers ensures that the model retains a high-level of feature richness throughout the processing pipeline, which allows for better preservation of fine-grained details, thereby enhancing the model's ability to capture the complex spatial relationships and subtle variations essential for distinguishing visually similar images. In addition, by avoiding intermediate size reductions, the model can prevent information loss that could otherwise hinder classification performance, particularly in unstructured natural settings devoid of clearly defined objects.

Furthermore, the experimental findings indicate that incorporating CNNs with attention processes improves the classification of images showing natural attractions. This corresponds with the findings of previous studies [40, 41] on attention-based models, e.g., the CBAM and SENet models, which utilize attention mechanisms to enhance the deep feature extraction of CNNs, thereby allowing the model to focus on relevant image regions more efficiently. Similarly, previous studies [32] have combined attention mechanisms with contextual information to enhance scene identification and address issues related to intraclass variances and interclass similarities.

Despite these advancements, the SENet model exhibits certain limitations, particularly in terms of optimizing the classification of highly similar natural scene images. The findings of this study indicate that its attention mechanisms are insufficiently adapted to handle images with substantial intraclass variance and interclass similarity effectively. To address these issues, this study has proposed the Attention-X method, which combines channel-wise features with deep feature maps derived from various CNN models, without distorting the spatial representation. This integration facilitates the identification of optimal configurations for various network topologies. In addition, the proposed Attention-X method proficiently categorizes images exhibiting significant intraclass variance and interclass similarity by utilizing channel attention with CNN deep features. As a result, the proposed method obtains performance that is similar to that of the CBAM model while eliminating the need for spatial attention or supplementary contextual data. The proposed model employs a multilevel attention mechanism, which enables it to focus on the most relevant features, thereby enhancing its capacity to manage interclass similarity and intraclass variance.

However, the class-wise F1-score comparison graph shown in Figure 8 reveals that the model continued to encounter challenges in accurately classifying certain scene categories, e.g., coasts, cliffs, hills, islets, mountains, oceans, and most notably valleys. The misclassification trend was uniform across all models assessed on the 16-challenge class dataset. The challenge in differentiating these classes arises from the existence of shared items and visually similar backgrounds, leading to overlapping feature representations. For example, beaches are similar to coasts, cliffs and caverns, and hills, mountains, and valleys exhibit similar forms and structural characteristics, thereby resulting in classification inaccuracies.

The limitations of the proposed method are demonstrated through feature visualization via scatter plots and the examination of regions emphasized by the model for classification using feature map heatmaps. This analysis focused on images from classes with low F1-scores, i.e., the hill and valley classes, using the Inception-V3-4X model. As shown in Figure 9, these examples highlight the model's ability to extract features from images prone to misclassification. When extracting features from inherently challenging classes using the Inception-V3-4X model, the heatmaps reveal the model's difficulty in focusing on distinguishing regions. These heatmaps emphasize regions of concern that inadequately distinguish between the two groups, resulting in classification problems. In addition, in both the images shown in Figures 9(a) and 9(b), most of the retrieved features have low activation values, as indicated by the dense aggregation of purple spots at the lower end of the activation distribution. This indicates that the model failed to adequately activate distinct discriminative features for hills, resulting in unclear classifications. The activation distributions for all classes are comparable, which suggests that they possess shared visual characteristics, including mountains, slopes, and vegetation. These findings indicate that future enhancements should concentrate on optimizing the attention mechanisms to increase the model's ability to differentiate visually similar natural environments.



(a) Inception-V3-4X vs. Inception-V3

(b) Inception-V3-4X vs. AlexNet

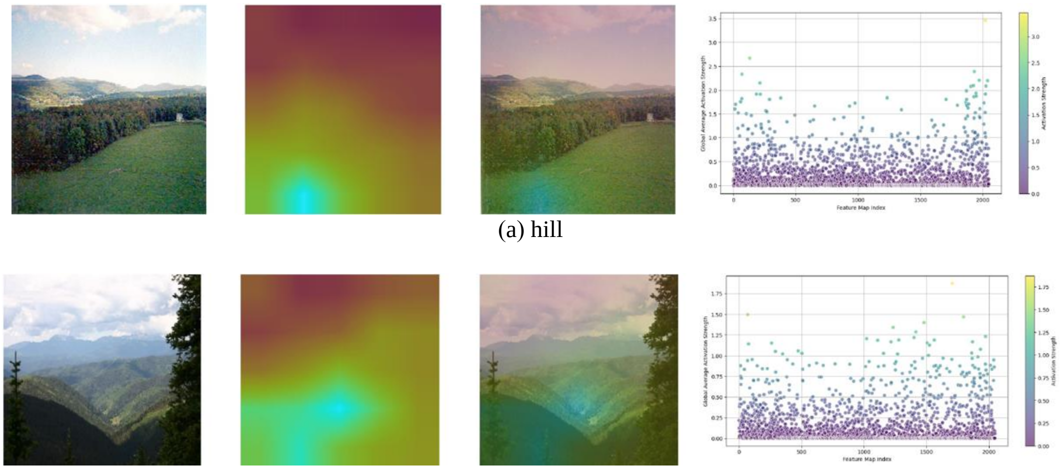
(c) Inception-V3-4X vs SAS

(d) Inception-V3-4X vs. Fine-tuned SAS

(e) Inception-V3-4X vs. CBAM

(f) Inception-V3-4X vs. SENet

Figure 8 Class-wise F1-score comparison with state-of-the-art models on 16 classes of the SUN397 dataset



(a) hill

(b) valley

Figure 9 Feature extraction and attention analysis for misclassified hill and valley classes

The investigation suggests that attention mechanisms require further refinement to enhance classification accuracy, particularly given the diverse and unpredictable characteristics of natural attraction scenes. Some images contain well-defined elements; however, others lack distinguishable features, exhibit cluttered backgrounds, or share similar color distributions across multiple classes, thereby complicating object recognition. This issue is particularly evident in images of valleys and hills with minimal grass coverage, where distinguishing between the visually similar features remains challenging. In addition, attention mechanisms enhance classification performance by emphasizing essential features and filtering out irrelevant information; however, this process may unintentionally disregard contextually significant information that, while not primary, still contributes to a comprehensive understanding of the scene. As a result, the model's ability to comprehend the broader context of the scene is diminished, particularly in cases where distinguishable objects or well-defined regions are lacking.

6. Conclusion and future work

The Inception-V3 model with 4X attention layers demonstrated the best overall accuracy, establishing it as the most precise model for natural scene classification. In addition, the ResNet-50 model with three attention layers exhibited a robust balance between accuracy and computational efficiency, rendering it the most effective model when evaluating both accuracy and complexity. This study further substantiates that the attention mechanism enables the model to focus on the most relevant regions of the image, resulting in improved classification outcomes. Integrating data from the original image and the emphasized attention regions allows for effective feature extraction, significantly enhancing the model's ability to distinguish visually similar scenes. Consequently, the proposed Attention-X significantly improves the efficacy of cutting-edge models in the classification of natural attraction scenes.

Nonetheless, despite the robust performance of the optimal model on the SUN397 and ADE20K datasets, the Places365 dataset continues to pose challenges because of its vastly diverse image content. This diversity adds complexity, hindering the model's ability to maintain consistently high accuracy throughout the dataset. Despite the effectiveness of the proposed Attention-X method in enhancing overall model accuracy, limitations remain—particularly in scenes characterized by complex backgrounds such as forests, hills, oceans, snowfields, valleys, and waterfalls. These scene categories often exhibit extensive background regions, similar color distributions, and large areas that visually resemble other classes, leading to reduced classification accuracy. This suggests that further refinement is necessary to improve differentiation among classes with overlapping or complex visual features.

Future work will focus on incorporating spatial attention mechanisms, which offer the potential to better identify relevant features based on spatial cues rather than relying solely on deep features from pretrained models. Additionally, multiscale attention maps will be explored to enhance feature extraction, especially in images containing objects of varying scales.

To address the common issue in CNNs, where spatial information is often lost due to pooling operations, semantic segmentation will be integrated into the attention mechanism. This integration aims to preserve spatial structure and ensure that attention is directed toward semantically meaningful regions, thereby improving the quality of learned representations and classification performance. Moreover, these enhancements are expected to extract complementary features that are especially valuable for scenes lacking distinct or discriminative objects.

To further assess the model's robustness, especially in handling intraclass variance and interclass similarity in outdoor recreational scenes—where both natural and man-made elements often coexist—future studies will also involve the exploration of additional scene categories. This would facilitate a more comprehensive evaluation of the model's effectiveness in real-world applications.

7. Acknowledgments

This paper is supported by the Faculty of Science, Naresuan University, and the Thailand Ministry of Higher Education, Science, Research, and Innovation.

8. References

[1] Cepeda-Pacheco JC, Domingo MC. Deep learning and internet of things for tourist attraction recommendations in smart cities. Neural Comput Appl. 2022;34(10):7691-709.

[2] Kitamura R, Itoh T. Tourist spot recommendation applying generic object recognition with travel photos. 22nd International Conference Information Visualisation (IV); 2018 Jul 10-13; Fisciano, Italy. USA: IEEE; 2018. p. 1-5.

[3] Katsumi H, Yamada W, Ochiai K. Characterizing generic POI: a novel approach for discovering tourist attractions. J Inf Process. 2023;31:265-77.

[4] Parikh V, Keskar M, Dharia D, Gotmare P. A tourist place recommendation and recognition system. 2018 Second International Conference on Inventive Communication and Computational Technologies; 2018 Apr 20-21; Coimbatore, India. USA: IEEE; 2018. p. 218-22.

[5] Sun S, Gong X. Hierarchical semantic contrast for scene-aware video anomaly detection. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17-24; Vancouver, Canada. USA: IEEE; 2023. p. 22846-56.

[6] Alqasrawi Y. Natural scene image annotation using local semantic concepts and spatial bag of visual words. Int J Sens Wirel Commun Control. 2016;6(3):153-73.

[7] Shahriari M, Bergevin R. Land-use scene classification: a comparative study on bag of visual word framework. Multimed Tools Appl. 2017;76(21):23059-75.

[8] Zhou Z, Li S, Wu W, Guo W, Li X, Xia G, et al. NaSC-TG2: Natural scene classification With Tiangong-2 remotely sensed imagery. IEEE J Sel Top Appl Earth Obs Remote Sens. 2021;14:3228-42.

[9] Gupta N, Khobragade P. Muti-class image classification using transfer learning. Int J Res Appl Sci Eng Technol. 2023;11(1):700-4.

[10] Sujee R, Sesh VB. Natural scene classification. 2019 International Conference on Computer Communication and Informatics; 2019 Jan 23-25; Coimbatore, India. USA: IEEE; 2019. p. 1-7.

[11] Xu C, Shu J, Wang Z, Wang J. A scene classification model based on global-local features and attention in lie group space. Remote Sens. 2024;16(13):2323.

[12] Liu Y, Zhong Y, Qin Q. Scene classification based on multiscale convolutional neural network. IEEE Trans Geosci Remote Sens. 2018;56(12):7109-21.

[13] Li J, Lin D, Wang Y, Xu G, Zhang Y, Ding C, et al. Deep discriminative representation learning with attention map for scene classification. Remote Sens. 2020;12(9):1366.

[14] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2006 Jun 17-22; New York, USA. USA: IEEE; 2006. p. 2169-78.

[15] Wilson J, Arif M. Scene recognition by combining local and global image descriptors [Internet]. arXiv [Preprint]. 2017 [cited 2024 Oct 30]. Available from: https://arxiv.org/abs/1702.06850.

[16] Xie L, Lee F, Liu L, Kotani K, Chen Q. Scene recognition: a comprehensive survey. Pattern Recognit. 2020;102:107205.

[17] Yang Y, Newsam S. Bag-of-visual-words and spatial extensions for land-use classification. Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems; 2018 Nov 6-9; Seattle, USA. San Jose California: ACM; 2010. p. 270-9.

[18] Singh S, Gupta A, Efros AA. Unsupervised discovery of mid-level discriminative patches [Internet]. arXiv [Preprint]. 2012 [cited 2024 Oct 30]. Available from: https://arxiv.org/abs/1205.3137.

[19] Sadeghi F, Tappen MF. Latent pyramidal regions for recognizing scenes. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C, editors. Computer Vision – ECCV 2012. Lecture Notes in Computer Science. Berlin: Springer; 2012. p. 228-41.

[20] Sitaula C, Shahi TB, Marzbanrad F, Aryal J. Recent advances in scene image representation and classification. Multimedia Tools Appl. 2024;83(3):9251-78.

[21] Ma Y, Lei Y, Wang T. A natural scene recognition learning based on label correlation. IEEE Trans Emerg Top Comput Intell. 2022;6(1):150-8.

[22] Yee PS, Lim KM, Lee CP. DeepScene: scene classification via convolutional neural network with spatial pyramid pooling. Expert Syst Appl. 2022;193:116382.

[23] Bai S, Tang H, An S. Coordinate CNNs and LSTMs to categorize scene images with multi-views and multi-levels of abstraction. Expert Syst Appl. 2019;120:298-309.

[24] Gao J, Yang J, Zhang J, Li M. Natural scene recognition based on convolutional neural networks and deep Boltzmannn machines. 2015 IEEE International Conference on Mechatronics and Automation; 2015 Aug 2-5; Beijing, China. USA: IEEE; 2015. p. 2369-74.

[25] Masood S, Ahsan U, Munawwar F, Rizvi DR, Ahmed M. Scene recognition from image using convolutional neural network. Procedia Comput Sci. 2020;167:1005-12.

[26] Sharma V, Nagpal N, Shandilya A, Dureja A, Dureja A. A practical approach to detect indoor and outdoor scene recognition. Proceedings of the 4th International Conference on Information Management & Machine Intelligence; 2022 Dec 23-24; Jaipur, India. New York: ACM; 2023. p. 1-10.

[27] Liu Y, Suen CY, Liu Y, Ding L. Scene classification using hierarchical wasserstein CNN. IEEE Trans Geosci Remote Sens. 2019;57(5):2494-509.

[28] Mungklachaiya S, Salaiwarakul A. Exploring deep learning features and bag-of-visual-words for scene classification. ICIC Express Lett B: Appl. 2024;15(10):1081-8.

[29] Baik S, Seong H, Lee Y, Kim E. Spatial-Channel transformer for scene recognition. 2022 International Joint Conference on Neural Networks; 2022 Jul 18-23; Padua, Italy. USA: IEEE; 2022. p. 1-8.

[30] Guo MH, Xu TX, Liu JJ, Liu ZN, Jiang PT, Mu TJ, et al. Attention mechanisms in computer vision: a survey. Comput Vis Media. 2022;8(3):331-68.

[31] Yang X. An overview of the attention mechanisms in computer vision. J Phys: Conf Ser. 2020;1693:012173.

[32] Peng Y, Liu X, Wang C, Xiao T, Li T. Fusing attention features and contextual information for scene recognition. Int J Pattern Recognit Artif Intell. 2022;36(3):2250014.

[33] Wang P, Qiao J, Liu N. An improved convolutional neural network-based scene image recognition method. Comput Intell Neurosci. 2022;2022(1):3464984.

[34] Liu R, Ning X, Cai W, Li G. Multiscale dense cross-attention mechanism with covariance pooling for hyperspectral image scene classification. Mob Inf Syst. 2021;2021(1):9962057.

[35] Zhang J, Yu X, Lei X, Wu C. A multi-feature fusion model based on denoising convolutional neural network and attention mechanism for image classification. Int J Swarm Intell Res. 2023;14(2):1-15.

[36] Ye W, Tan R, Liu Y, Chang CC. The comparison of attention mechanisms with different embedding modes for performance improvement of fine-grained classification. IEICE Trans Inf Syst. 2023;E106.D(5):590-600.

[37] Xiao J, Ehinger KA, Hays J, Torralba A, Oliva A. SUN Database: exploring a large collection of scene categories. Int J Comput Vis. 2016;119(1):3-22.

[38] Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A. Places: a 10 million image database for scene recognition. IEEE Trans Pattern Anal Mach Intell. 2018;40(6):1452-64.

[39] López-Cifuentes A, Escudero-Viñolo M, Bescós J, García-Martín Á. Semantic-aware scene recognition. Pattern Recognit. 2020;102:107256.

[40] Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. Computer Vision – ECCV 2018: 15th European Conference; 2018 Sep 8-14; Munich, Germany. Berlin: Springer; 2018. p. 3-19.

[41] Hu J, Shen L, Sun G. Squeeze-and-Excitation networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18-23; Salt Lake City, USA. USA: IEEE; 2018. p. 7132-41.

[42] Zhou B, Zhao H, Puig X, Fidler S, Barriuso A, Torralba A. Scene parsing through ADE20K dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21-26; Honolulu, USA. USA: IEEE; 2017. p. 5122-30.