Channel and Spatial attention based CNN model for Texture image analysis

MINOR PROJECT-2 REPORT

Submitted by

ASHOK.Y

SRIKANTH.M

PRASAD.N

Under the Guidance of

Dr.V.GNANAPRAKASH

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

ELECTRONICS & COMMUNICATION ENGINEERING

Vel Tech
Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology
(Deemed to be University Estd. u/s 3 of UGC Act, 1956)

OCTOBER 2024

Vel Tech
Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology
(Deemed to be University Estd. u/s 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that this Minor project-2 report entitled**CHANNEL AND SPATIAL ATTENTION BASED CNN MODEL FOR TEXTURE IMAGE ANALYSIS** is the bonafide work of **Y.ASHOK(21UEEA0139),M.SRIKANTH (21UEEA0198) and N.PRASAD (21UEEA0202)** who carried out the project work under my supervision.

SUPERVISOR

Dr.V.GNANAPRAKASH
Assistant Professor
Department of ECE

HEAD OF THE DEPARTMENT

Dr.A. SELWIN MICH PRIYADHARSON
Professor
Department of ECE

Submitted for Minor project-2 work viva-voce examination held on:_____

INTERNAL EXAMINER

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

We express our deepest gratitude to our Respected Founder President and Chancellor **Col. Prof. Dr. R. Rangarajan**, Foundress President **Dr. R. Sagunthala Rangarajan**, Chairperson and Managing Trustee and Vice President.

We are very thankful to our beloved Vice Chancellor **Prof. Dr. Rajat Gupta** for providing us with an environment to complete the work successfully.

We are obligated to our beloved Registrar **Prof.Dr. E. Kannan** for providing immense support in all our endeavours. We are thankful to our esteemed Dean Academics **Prof.Dr.Raju Shanmugam** for providing a wonderful environment to complete our work successfully.

We are extremely thankful and pay my gratitude to our Dean SoEC **Prof.Dr.R.S.Valarmathi** for her valuable guidance and support on completion of this project.

It is a great pleasure for us to acknowledge the assistance and contributions of our Head of the De- partment **Prof.Dr.A.Selwin Mich Priyadharson**, Professor for his useful suggestions, which helped us in completing the work in time and we thank him for being instrumental in the completion of third year with his encouragement and unwavering support during the entire course. We are extremely thankful and pay our gratitude to our Minor Project-2 coordinator **Dr.Hari krishnana paik** , for his valuable guidance and support in completing this project report in a successful manner..

We are grateful to our supervisor **Dr.JEEN RETNA KUMAR**, Assistant Professor ECE for his guidance and valuable suggestion to carry out our project work successfully.

We thank our department faculty, supporting staffs and our family and friends for encouraging and supporting us throughout the project.

HEMANTH NAIDU.Y

JAYA SEKHAR. D

DATTA SAI TEJA. A

TABLE OF CONTENTS

ABSTRACT

The accurate classification of architectural styles is of great significance to the study of architectural culture and human historical civilization. Models based on convolutional neural network (CNN) have achieved highly competitive results in the field of architectural image classification owing to its more powerful capability of feature expression.CNN models to date only extract the global features of architecture facade or focus on some regions of architecture and fail to extract the spatial features of different components. To improve the accuracy of architectural style classification, we propose an architectural image classification method based on CNN and channel–spatial attention. Firstly, we add a preprocessing operation before CNN feature extraction to select main building candidate region in architectural image and then use CNN feature extractor for deep feature extraction. Secondly, channel–spatial attention module is introduced to generate an attention map, which can not only enhance the texture feature representation of architectural images but also focus on the spatial features of different architectural elements.

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Convolutional block attention module (CBAM)

This project investigates the role of Convolutional Neural Network (CNN) architectures in binary image classification, specifically for distinguishing between images of cats and dogs. The models used—VGG-16, VGG-19, and Res Net-50—are well-known for their ability to capture and process complex visual patterns. However, traditional CNNs may lack the ability to focus on the most relevant features within an image, which can limit performance in fine-grained classification tasks.

CNNs are a class of deep neural networks widely used for image classification, segmentation, and other image-related tasks. For texture image analysis, CNNs can automatically learn spatial hierarchies of features, which is crucial for understanding complex texture patterns. The network extracts features at various levels of abstraction, starting from low-level features (edges, corners) to higher-level patterns (texture patches). Attention mechanisms allow a model to dynamically focus on specific parts of the input data, making the network more flexible and efficient.

The results showed that ResNet-50 without CBAM achieved the highest validation accuracy (98.43The goal of the project was to classify input images of size (224 x 224 x 3) into two classes: '0' for cats and '1' for dogs. Cross-entropy loss was used as the loss function, and several hyperparameters such as the number of epochs, learning rate, momentum, weight decay, and batch size were tuned for optimal performance. The dataset, which is split into training, validation, and testing sets, was used to evaluate the performance of each model by comparing training accuracy and validation accuracy.

Apart from these factors, we investigate a different aspect of the architecture design, attention. The significance of attention has been studied extensively in the previous literature [12–17]. Attention not only tells where to focus, it also improves the representation of interests. Our goal is to increase representation power by using attention mechanism: focusing on important features and sup- pressing unnecessary ones. In this paper, we propose a new network module, named "Con-

volutional Block Attention Module". Since convolution operations extract informative features by blending cross-channel and spatial information together, we adopt our module to emphasize meaningful features along those two principal dimensions: channel and spatial axes. To achieve this, we sequentially apply channel and spatial attention modules (as shown in Fig. 1), so that each of the branches can learn 'what' and 'where' to attend in the channel and spatial axes respectively. As a result, our module efficiently helps the information flow within the network by learning which information to emphasize or suppress. In the ImageNet-1K dataset, we obtain accuracy improvement from various baseline networks by plugging our tiny module, revealing the efficacy of CBAM.We visualize trained models using the grad-CAM [18] and observe that CBAM enhanced networks focus on target objects more properly than their baseline networks. Taking this into account, we conjecture that the performance boost comes from accurate attention and noise reduction of irrelevant clutters. Finally, we validate performance improvement of object detection on the MS COCO and the VOC 2007 datasets, demonstrating a wide applicability of CBAM. Since we have carefully designed our module to be light-weight, the overhead of parame- ters and computation is negligible in most case.

1.2 Channel- Spatial attention Module:

The CBAM module operates by refining the feature maps generated by CNN layers through two types of attention: channel attention (which highlights the most relevant channels) and spatial attention (which focuses on important spatial regions). These modifications allow the model to prioritize the most important features, potentially improving classification performance.

Channel and spatial attention-based CNN (Convolutional Neural Network) models have shown remarkable success in a variety of image processing tasks, including texture image analysis. The integration of attention mechanisms enhances the model's ability to focus on important features by adaptively adjusting the importance of different regions or channels in the input data. Here's a breakdown of how channel and spatial attention mechanisms work within a CNN for texture image analysis. Convolutional Neural Networks (CNNs) for Texture Image Analysis.

In comparing the performance of these models, both with and without CBAM, the project demonstrates how attention mechanisms can potentially enhance model performance, highlighting the importance of feature selection in CNN architectures for image classification tasks. The ultimate goal is to measure the improvement in accuracy and generalization ability, providing valuable insights into the utility of attention modules in deep learning models.

To address this, we integrate the Convolutional Block Attention Module (CBAM) into these architectures. CBAM enhances feature representation by applying spatial and channel attention mechanisms, which direct the network's focus to the most informative parts of the image while filtering

out less relevant details. This project compares models with and without CBAM to understand how attention mechanisms affect classification performance.

Each model, pre-trained on Image Net, is fine-tuned on the cat-dog dataset, with input images resized to 224 x 224 pixels. The models are trained using cross-entropy loss, and hyperparameters like learning rate, momentum, weight decay, and batch size are optimized to improve performance. By analyzing training and validation accuracy, as well as generalization on test data, this study aims to provide insights into the effectiveness of attention mechanisms for CNN-based image classification.

1.3 CNN architectures like VGG-16, VGG-19, and ResNet-50

CNN architectures like VGG-16, VGG-19, and ResNet-50 have become central to computer vision, excelling in capturing hierarchical features through deep, layered structures. However, while these architectures have shown exceptional results, they sometimes struggle to differentiate between fine details within similar classes, such as cats and dogs. This limitation can be partly attributed to their inability to selectively focus on the most relevant features within an image.

To address this challenge, we incorporate the Convolutional Block Attention Module (CBAM) into these CNN models. CBAM introduces two levels of attention: channel attention, which allows the model to prioritize specific channels, and spatial attention, which enhances important spatial regions. Together, these mechanisms help the model selectively emphasize informative features and ignore irrelevant areas, potentially improving classification accuracy and robustness.

The project utilizes models pre-trained on the ImageNet dataset, fine-tuning them on a cat-and-dog image dataset. The training process uses cross-entropy loss, and hyperparameters—such as learning rate, momentum, weight decay, and batch size—are optimized for best performance. Each model's performance is evaluated by comparing training and validation accuracy, with and without the CBAM module. Through these experiments, the project aims to reveal the impact of attention mechanisms on feature extraction and classification accuracy, highlighting the potential of attention modules to enhance CNN performance in image classification tasks.

1.4 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) have become the standard approach for image classification tasks due to their ability to automatically learn hierarchical features from raw images. Early architectures such as LeNet-5 and AlexNet paved the way for modern, deeper models like VGG and ResNet, which have demonstrated state-of-the-art performance on large-scale datasets such as ImageNet. VGG-16 and VGG-19 introduced a simple yet effective architecture that stacks convolutional layers, but their depth makes them computationally expensive. On the other hand, ResNet-50, with

its residual connections, addresses the vanishing gradient problem in deep networks, enabling significantly deeper architectures to be trained.

To further enhance the performance of CNNs, recent work has introduced attention mechanisms that allow models to focus on the most relevant parts of an image. One such development is the Convolutional Block Attention Module (CBAM), proposed by Woo et al. (2018), which applies attention at both the spatial and channel levels. This mechanism helps CNNs refine feature maps, allowing them to prioritize important features and ignore less relevant information, thereby improving classification accuracy. Studies show that integrating attention mechanisms like CBAM into CNN architectures can lead to improved feature representation and, in some cases, enhanced classification performance, particularly in challenging tasks where subtle distinctions between classes exist.

Several works have compared CNN architectures with and without attention mechanisms for various tasks. For example, CBAM has been applied successfully to tasks like object detection, semantic segmentation, and image classification, showing noticeable improvements in the network's ability to generalize to unseen data. However, the extent of improvement may vary depending on the task, dataset, and underlying model architecture.

We collected a dataset of dog and cat images for image classification. After collecting the data, we applied the Convolutional Block Attention Module (CBAM) to improve the model's representation power. CBAM uses an attention mechanism that focuses on important features while suppressing unnecessary ones. It consists of two sequential sub-modules: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM).

Next, we split the dataset into training, testing, and validation sets to evaluate model performance. During training, we utilized CNN architectures, including VGG-16, VGG-19, and ResNet-50, to achieve accuracy for the models. In this project, different CNN Architectures like VGG-16, VGG-19, and ResNet-50, were used for the task of Dog-Cat image classification. The input to the CNN networks was a (224 x 224 x 3) image and the number of classes were 2, where '0' was for a cat and '1' was for a dog.

In this project, we build upon this existing body of work by comparing CNN architectures—VGG-16, VGG-19, and ResNet-50—with and without CBAM for the task of cat-dog image classification. The goal is to assess how spatial and channel attention mechanisms influence feature extraction and classification accuracy, providing insights into the effectiveness of attention modules in real-world image classification scenarios.

CHAPTER 2

LITERTAURE SURVEY

2.1 Overview

Recent advancements in computer vision have positioned Convolutional Neural Networks (CNNs) as the primary approach for image classification tasks, due to their ability to learn complex, hierarchical patterns in image data. Landmark architectures, such as Le Net and AlexNet, introduced CNNs as powerful tools for visual recognition. Building upon this foundation, deeper networks like VGG and ResNet have demonstrated state-of-the-art performance on large-scale datasets, with VGG models utilizing deep stacks of convolutional layers for feature extraction and ResNet using residual connections to mitigate vanishing gradient issues in deeper networks. Despite their success, traditional CNNs often lack the ability to focus on the most relevant regions within an image, which can limit their performance, particularly in tasks with subtle distinctions between classes. To address this, attention mechanisms have emerged as a method to direct the network's focus to critical parts of the image. The Convolutional Block Attention Module (CBAM), proposed by Woo et al. (2018), is one such module that applies spatial and channel attention, helping models to dynamically weigh the importance of different feature maps and spatial regions.

Studies have shown that integrating CBAM or similar attention mechanisms into CNN architectures can lead to improved performance in tasks such as object detection, segmentation, and image classification. For instance, CBAM has demonstrated efficacy in refining feature representations, allowing models to generalize better on test data and achieve higher accuracy. This project builds upon these findings by integrating CBAM into popular CNN architectures (VGG-16, VGG-19, and ResNet-50) for the specific task of classifying cats and dogs. By evaluating models both with and without CBAM, the project aims to assess the impact of attention mechanisms on feature selection and classification accuracy, contributing to a deeper understanding of how attention modules can enhance CNN performance in real-world image classification applications.

2.2 Litertaure survey

2.2.1 Architectural style classification based on CNN and channel-spatial attention (Sulan Zhang, 2022)

Journal: Computer Science and Technology Description: This study presents a deep learning approach focused on classifying architectural styles. The author used a CNN model enhanced by channel-spatial attention mechanisms. This combination enables the model to better capture and focus on key features of architectural styles, improving the classification accuracy.

2.2.2 SSTNet: Spatial, Spectral, and Texture Aware Attention Network using Hyperspectral Image for Corn Variety Identification (Weidong Zhang and Zexu Li, 2019)

Journal: School of Information Engineering Description: The authors propose SSTNet, an attention network that integrates spatial, spectral, and texture-aware attention mechanisms. By using hyperspectral imaging, the network can more accurately identify different varieties of corn based on subtle differences in their spectral signatures and textures. This model significantly enhances the performance in agricultural image classification tasks.

2.2.3 SCA-Net: A Spatial and Channel Attention Network for Medical Image Segmentation (Tong Shan and Jiayong Yan, 2021)

Journal: Health Science and Engineering Description: This research focuses on the SCA-Net, which combines spatial and channel attention mechanisms to improve segmentation accuracy in medical images. The model enhances the network's ability to focus on crucial areas in medical images, leading to better segmentation results. This is particularly beneficial in diagnostic and treatment planning applications, where precision in image analysis is critical.

2.2.4 Combined Channel and Spatial Attention-based Stereo Endoscopic Image Super-Resolution (Dr. Titipat Achakulvisut, 2020)

Journal: Department of Biomedical Engineering Description: This study introduces the use of channel and spatial attention in stereo endoscopic imaging to improve the resolution of medical images. The integration of these attention mechanisms brings a significant revolution in medical diagnostics and surgeries, allowing for better image clarity and precision during procedures.

2.2.5 CNN Cloud Detection Algorithm Based on Channel and Spatial Attention and Probabilistic Upsampling for Remote Sensing Image (Jing Zhang, 2021)

Journal: PhD degree in Information and Communication Engineering Description: This research presents a CNN-based cloud detection algorithm for remote sensing images, enhanced by channel and spatial attention mechanisms and probabilistic upsampling. The algorithm aims to identify and remove irrelevant cloud pixels, which often interfere with remote sensing images. By improving cloud detection accuracy, the method reduces the transmission of unnecessary data and improves the overall quality of the information captured.

2.2.6 A Study on Super-Resolution Using Spatial and Channel Attention (Dongwoo Lee, 2023)

Dongwoo Lee introduces an approach for single-image super-resolution by incorporating spatial and channel attention mechanisms. This model emphasizes high-frequency features, which are essential in tasks like super-resolution where fine details are crucial. By using spatial attention to focus on important image regions and channel attention to refine feature selection, the proposed model achieves improved image resolution and quality. This study highlights the importance of attention mechanisms in enhancing image details, particularly in applications requiring fine-grained feature emphasis.

2.2.7 CCNet: CNN Model with Channel Attention and Convolutional Pooling for Spatial Image Steganalysis (Tong Fu Liquan Chen, 2022)

In this study, Tong Fu and Liquan Chen propose CCNet, a CNN-based model designed for image steganalysis. This model uses a channel attention mechanism to focus on regions suitable for embedding hidden information, a key aspect in the detection of steganography. The channel attention mechanism enables the network to adaptively select important features, enhancing its ability to identify subtle patterns or artifacts left by embedded information. The study demonstrates the utility of channel attention in tasks requiring high sensitivity to subtle changes in spatial data, such as steganalysis.

2.2.8 TSCA-Net: Transformer-Based Spatial-Channel Attention Segmentation Network for Medical Images (Yinghua Fu, 2022)

Yinghua Fu presents TSCA-Net, an encoder-decoder model that uses a Transformer-based approach for spatial and channel attention. Designed specifically for medical image segmentation, this model combines the strengths of both CNNs and Transformers, allowing it to capture fine details in medical images, which is crucial for accurate segmentation. Spatial attention enables the model to localize important regions, while channel attention focuses on relevant feature maps. The study illustrates how attention mechanisms can enhance segmentation accuracy in high-stakes applications, such as medical imaging, where precise feature extraction and localization are vital. These studies underscore the versatility and impact of spatial and channel attention mechanisms in enhancing CNN and Transformer-based models across different domains. By refining feature selection and emphasizing crucial details, attention mechanisms contribute to improved model performance in tasks ranging from image resolution enhancement to medical image segmentation and steganalysis. This project builds on these insights by exploring the impact of Convolutional Block Attention Module (CBAM), which combines spatial and channel attention, on CNN architectures (VGG-16, VGG-19, ResNet-50) for the binary image classification task of distinguishing between cats and dogs.

CHAPTER 3

METHODOLOGY OF CNN AND CHANNEL-SPATIAL ATTENTION

3.1 Proposed method:

To solve the problem of missing spatial features of different architecture components when extracting architectural fea- tures, we propose an architectural image classification method based on CNN and channel–spatial attention. our method is mainly composed of four parts: selective preprocessing module, CNN feature extractor, CSAM and classifier learning. Firstly, we use the selective preprocessing module to obtain the main candidate area of the ancient building. Secondly, the candidate region is used as the input of the feature extractor for deep feature extraction. Then, we introduce CSAM to focus on the texture features of building images and the spatial features of different architecture elements. Finally, the Soft max classifier is utilized to predict the score of the target class

3.1.1 Selective preprocessing module:

The goal of the selective preprocessing module is to select candidate regions of the main building from the image, including selection search algorithm and CASC. The selective search algorithm has the characteristics of diversified sampling techniques due to the combination of the advantages of exhaustive search and segmentation, thus allowing the generation of possible object locations for object recognition. Therefore, we introduce this algorithm to capture all object candidate regions of building images. Specifically, a graph-based segmentation method is first used to obtain pixel-level segmented regions, and then, the regions are merged according to the multiple similarity strategy from the selective search algorithm to capture the candidate regions of all objects on the building image. Since architectural object occupies large region in architec- tural image, we use the CASC method to calculate the number of pixels in the candidate regions of all objects and select the region with the largest number of pixels. The entire selective preprocessing algorithm process is shown in Algorithm.

3.1.2 Block Diagram

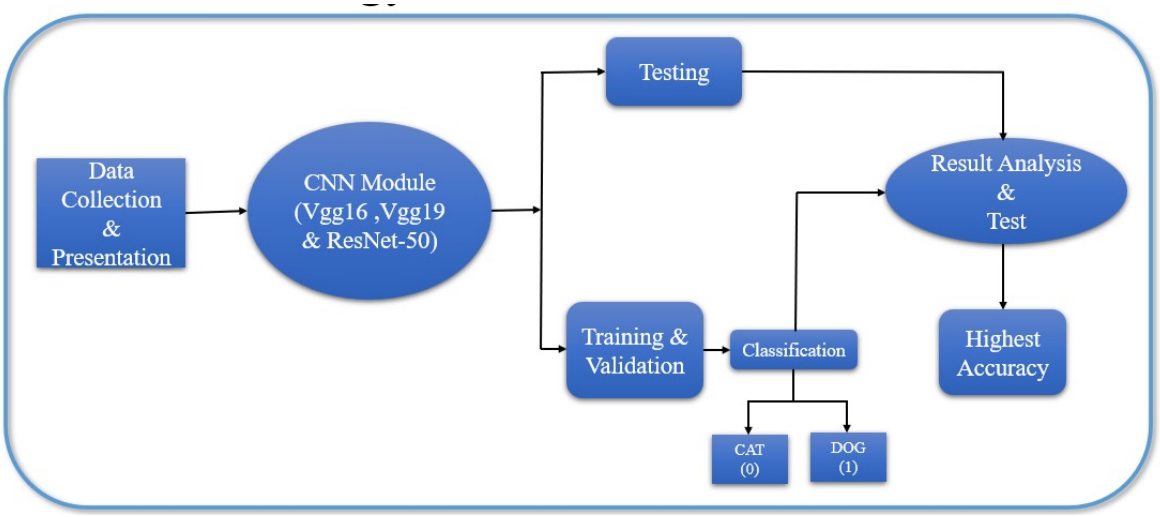

Figure 3.1: CNN methodology

3.1.3 Data Collection Presentation:

This step involves gathering the dataset of cat and dog images, likely including preprocessing tasks such as resizing, normalization, and splitting into training, validation, and test sets. Proper data presentation helps improve model performance and consistency during training.

3.1.4 CNN Module (VGG-16, VGG-19, ResNet-50):

The selected CNN architectures (VGG-16, VGG-19, ResNet-50) are trained on the prepared dataset. These architectures are pre-trained on ImageNet but can be fine-tuned for the specific task of classifying cats and dogs. This module may also include the Convolutional Block Attention Module (CBAM) if attention mechanisms are tested.

3.1.5 Training Validation:

The CNN models are trained on the training dataset, and validation data is used to evaluate the model's generalization ability during training. This helps in tuning hyper parameters and preventing overfitting.

3.1.6 Classification:

The trained model performs classification, where it assigns a label, either cat (0) or dog (1), to each image. This classification process is the main task, as the model learns to differentiate between

the two classes based on the features it has extracted.

3.1.7 Testing:

Once training and validation are complete, the model's performance is evaluated on the test dataset. Testing provides an unbiased measure of how well the model performs on unseen data, simulating real-world use.

3.1.8 Result Analysis Test:

The results are analyzed to assess the model's accuracy and performance metrics on both validation and test sets. This stage may also include comparison across different models or configurations (e.g., with and without CBAM).

3.1.9 Highest Accuracy:

After testing and analysis, the best-performing model configuration is identified based on the highest accuracy or other performance metrics. This model is deemed the most effective for the classification task.

This diagram reflects a comprehensive process for training, validating, and testing CNN models with the goal of achieving optimal performance on a binary classification task.

CNN Algorithm:

A Convolutional Neural Network (CNN) is a type of Deep Learning neural network architecture commonly used in Computer Vision. Convolutional Neural Network (CNN) is the extended version of artificial neural networks (ANN) which is predominantly used to extract the feature from the grid-like matrix dataset.VGG16, VGG19, and ResNet-50 are popular convolutional neural network (CNN) architectures.

1. VGG16 (Visual Geometry Group 16) Layers: 16 weight layers (13 convolutional layers and 3 fully connected layers). - Key Feature: Uses small (3 times 3) filters consistently throughout the network. It stacks multiple convolution layers before max-pooling layers, which helps in capturing detailed features. - Performance: High accuracy but requires a large amount of computation and memory.

2. VGG19 - Layers: 19 weight layers (16 convolutional layers and 3 fully connected layers). Key Feature: Similar to VGG16, but with more convolutional layers. The increased depth can capture

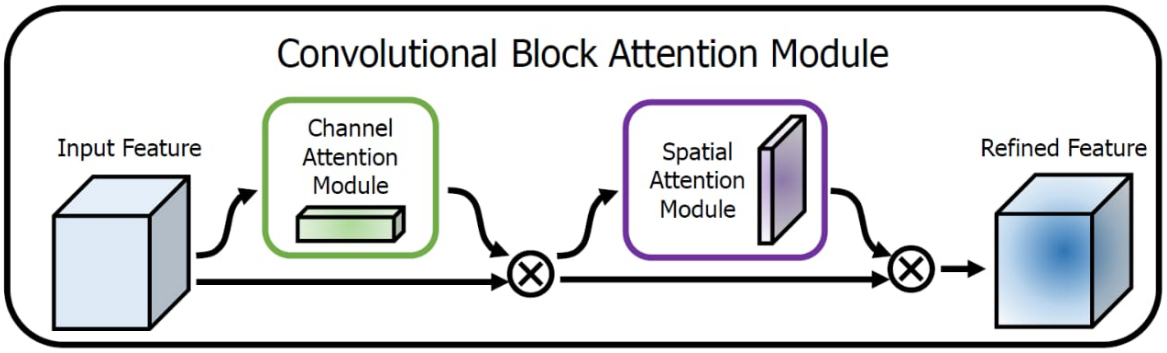

Figure 3.2: Convolutional block attention module

more complex features, but this also increases the model size and training time. - Performance: Slightly higher accuracy than VGG16 but with similar challenges in terms of resource usage.

3. ResNet-50 (Residual Network) - Layers: 50 layers with residual connections. - Key Feature: Introduces skip connections or residual connections, which help solve the vanishing gradient problem by allowing gradients to flow directly through the network. - Performance: Deep architecture with significantly fewer parameters compared to VGG networks and higher accuracy, thanks to the efficient use of residual blocks.

3.2 CNN feature extractor:

The feature map of the building can be obtained by using the CNN feature extractor to perform feature aggregation on the candidate area. We use the Inception-v3 model [3.2] for fea- ture extraction. The Inception-v3 model consists of several modules connected in a certain order. On the one hand, these modules are composed of convolutional layers and pooling layers of different scales, which can extract context informa- tion at multiple scales. On the other hand, these convolution and pooling operations can be performed in parallel, which makes the Inception-v3 model more effective than other CNN models.

3.3 Channel–spatial attention module:

CSAM is mainly composed of two parts.channel attention operation and spatial attention operation. They learn the what and where of attention, respectively. Recalling the Convolutional Block Attention Module (CBAM), it is mainly composed of two parts: channel attention module (CAM) and spatial attention module(SAM). Specifically, the structure of the attention.after the channel attention operation will get the attention map. Then, the resulting attention.

channel attention module:

A Channel Attention Module is a module for channel-based attention in convolutional neural networks. We produce a channel attention map by exploiting the inter-channel relationship of features. As each channel of a feature map is considered as a feature detector, channel attention focuses on 'what' is meaningful given an input image. To compute the channel attention efficiently, we squeeze the spatial dimension of the input feature map.

We follow the principle of the channel attention module in CBAM, and average-pooling and max-pooling are employed to aggregate the channel information of the generated build- ing feature map, generating two different channel attention descriptors: Fcavg and Fcmax. The channel attention map Mc RC×1×1 is generated by inputting two descriptors to a shared network. The shared network consists of a hidden layer multilayer perceptron (MLP).

Channel attention module. We produce a channel attention map by exploit- ing the inter-channel relationship of features. As each channel of a feature map is considered as a feature detector [31], channel attention focuses on 'what' is meaningful given an input image. To compute the channel attention efficiently,we squeeze the spatial dimension of the input feature map. For aggregating spatial information, average-pooling has been commonly adopted so far. Zhou et al.suggest to use it to learn the extent of the target object effectively and Hu eqal. adopt it in their attention module to compute spatial statistics. Beyond the previous works, we argue that max-pooling gathers another important clue about distinctive object features to infer finer channel-wise attention. Thus, we use both average-pooled and max-pooled features simultaneously. We empirically confirmed that exploiting both features greatly improves representation power of networks rather than using each independently (figure 3.4), showing the effectiveness of our design choice.

Diagram of each attention sub-module. As illustrated, the channel sub-module utilizes both max-pooling outputs and average-pooling outputs with a shared network; the spatial sub-module utilizes similar two outputs that are pooled along the channel axis and forward them to a convolution layer. We first aggregate spatial information of a feature map by using both average- pooling and max-pooling operations, generating two different spatial context descriptors: Fcavg and Fcmax, which denote average-pooled features and max-pooled features respectively. Both descriptors are then forwarded to a shared network to produce our channel attention map Mc RC×1×1.The shared network is composed of multi-layer perceptron (MLP) with one hidden layer. To reduce parameter overhead, the hidden activation size is set to RC/r×1×1,where r is the reduction ratio. After the shared network is applied to each descriptor, wemerge the output feature vectors using element-wise summation. In short, the channel attention is computed as:

$$\mathrm{Mc(F)} = (\ (\mathrm{MLP(AvgP\ ool(F))} + \mathrm{MLP(Maxool(F))})) = (\mathrm{W1(W0(Fcavg))} + \mathrm{W1(W0(Fcmax))}),$$

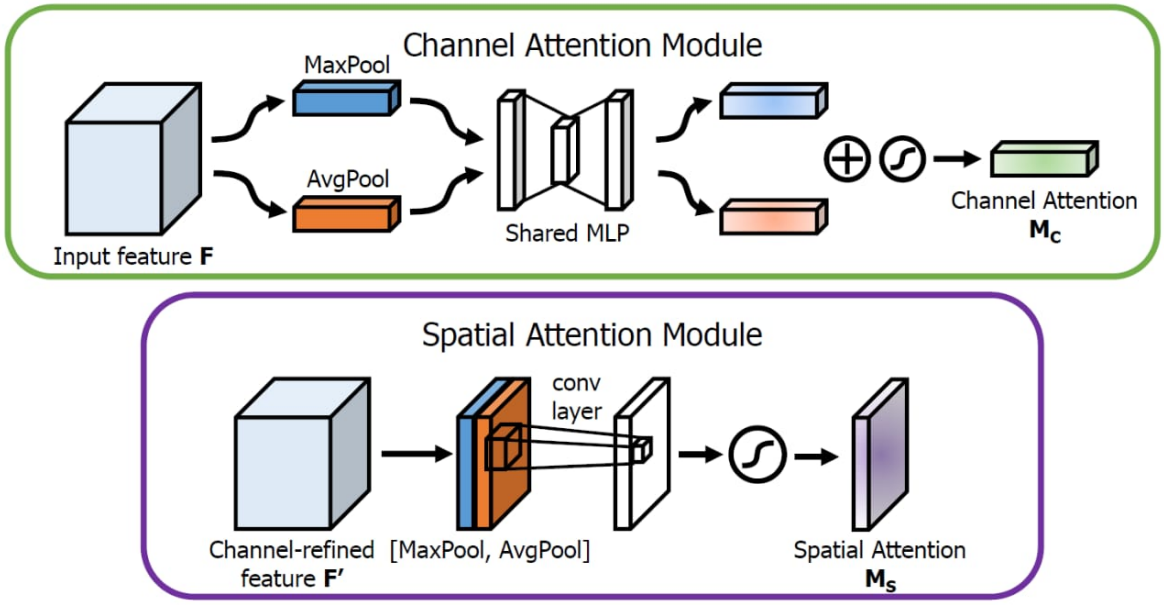

Figure 3.3: Channel-Spatial attention module

where denotes the sigmoid function, W0 RC/r×C , and W1 RC×C/r. Note that the MLP weights, W0 and W1, are shared for both inputs and the ReLU activation function is followed by W0.

Spatial attention module:

A Spatial Attention Module is a module for spatial attention in convolutional neural networks. It generates a spatial attention map by utilizing the inter-spatial relationship of features. Different from the channel attention, the spatial attention focuses on where is an informative part, which is complementary to the channel attention. To compute the spatial attention, we first apply average-pooling and max-pooling operations along the channel axis and concatenate them to generate an efficient feature descriptor.

To extract the spatial information of different components of the building, we use the spatial relationship between the features to generate a spatial attention map. The information of spatial features is focused on spatial attention operation,which is a supplement to channel attention. Average pooling and max-pooling operations are applied along the channel layers, and concatenated feature descriptors are generated by connecting both operations. Then, the concatenated feature descriptors are applied to a convolution operation for obtaining the spatial attention map MS RH×W which encodes location to emphasize or suppress. Specifically, average pooling and max-pooling will generate two 2D maps.

We generate a spatial attention map by utilizing the inter-spatial relationship of features. Different from the channel attention,the spatial attention focuses on ‘where’ is an informative part,

which is complementary to the channel attention. To compute the spatial attention, we first apply average-pooling and max-pooling operations along the channel axis and concatenate them to generate an efficient feature descriptor. Applying pooling operations along the channel axis is shown to be effective in highlighting informative regions . On the concatenated feature descriptor, we apply a convolution cnn of image analysis cats and dogs.

We collected a dataset of dog and cat images for image classification. After collecting the data, we applied the Convolutional Block Attention Module (CBAM) to improve the model's representation power. CBAM uses an attention mechanism that focuses on important features while suppressing unnecessary ones. It consists of two sequential sub-modules: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM).

Next, we split the dataset into training, testing, and validation sets to evaluate model performance. During training, we utilized CNN architectures, including VGG-16, VGG-19, and ResNet-50, to achieve accuracy for the models. In this project, different CNN Architectures like VGG-16, VGG-19, and ResNet-50, were used for the task of Dog-Cat image classification. The input to the CNN networks was a (224 x 224 x 3) image and the number of classes were 2, where '0' was for a cat and '1' was for a dog.

Arrangement of attention modules:

Given an input image, two attention modules, channel and spatial, compute complementary attention, focusing on 'what' and 'where' respectively. Considering this, two modules can be placed in a parallel or sequential manner. We found that the sequential arrangement gives a better result than a parallel arrangement. For the arrangement of the sequential process, our experimental result shows that the channel-first order is slightly better than the spatial-first. We will discuss experimental results on network engineering.

Comparison of different spatial attention methods. Using the proposed channel-pooling

we compare three different ways of arranging the channel and spatial attention submodules: sequential channel-spatial, sequential spatial-channel, and parallel use of both attention modules. As each module has different functions, the order may affect the overall performance. For example, from a spatial viewpoint, the channel attention is globally applied, while the spatial attention works locally. Also, it is natural to think that we may combine two attention outputs to build a 3D attention map. In the case, both attentions can be applied in parallel, then the outputs of the two attention modules are added and normalized with the sigmoid function.

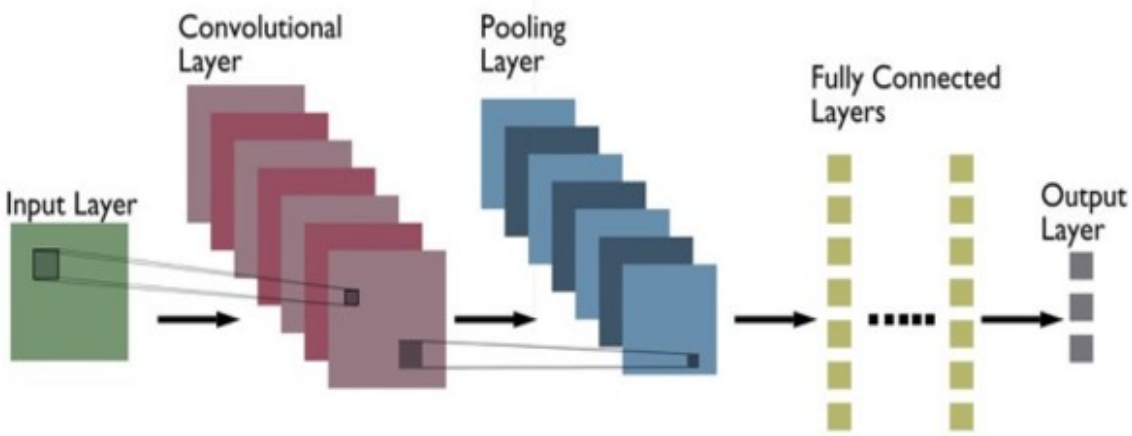

Figure 3.4: CNN layers

3.3.1 Input Layer:

This is where the raw image data is fed into the network. The image data is represented in pixel values, typically with dimensions like (224 x 224 x 3) for an RGB image, where 3 indicates the color channels (Red, Green, and Blue).

3.3.2 Convolutional Layer:

The convolutional layer applies filters to the input image to extract various features, such as edges, textures, and patterns. Each filter slides over the input image, performing element-wise multiplications and summing the results to produce a feature map. Multiple filters allow the network to capture a range of visual features at different locations.

3.3.3 Fully Connected Layers:

These layers take the flattened feature maps from the previous layers and connect them to the neurons in a fully connected fashion. They combine features learned in earlier layers to make predictions. The final fully connected layer typically outputs a vector with probabilities for each class.

3.3.4 Output Layer:

The output layer provides the final prediction, where each neuron corresponds to a specific class (e.g., cat or dog in a binary classification task). The highest value in this layer represents the predicted class.

This structure demonstrates a standard flow in CNNs: from feature extraction in the convolutional and pooling layers to classification in the fully connected layers. This architecture is commonly used in image recognition and classification tasks.

CHAPTER 4

CONCLUSION

The ResNet-50 architecture outperforms both VGG-16 and VGG-19 in terms of validation accuracy, This suggests that ResNet-50 is better at generalizing to unseen data for the Dog-Cat image classification task.Adding CBAM (Convolutional Block Attention Module), which introduces spatial and channel attention mechanisms, could potentially improve the model's performance further by helping the network focus on more important parts of the images. However, results for architectures with CBAM were not provided in this specific summary.VGG-19 has better validation accuracy than VGG-16, though both architectures achieve high training and validation accuracies, indicating they perform well but may not generalize as effectively as ResNet-50.ResNet-50's deeper architecture and residual connections likely contribute to its higher performance compared to the VGG models.

Visual Geometric group(VGG-16)

```
    print()
    print("Epoch" + str(epoch) + ":")
    print("Training Accuracy: " + str(training_accuracy) + "    Validation Accuracy: " + str(valid_accuracy))
    print("Training Loss: " + str(training_loss) + "    Validation Loss: " + str(valid_loss))
    print()
```

```
Epoch0:
Training Accuracy: 97.62889879057924    Validation Accuracy: 98.16200390370852
Training Loss: 0.060913320690377665    Validation Loss: 0.04680526513677788
```

Figure 4.1: VGG-16

Visual Geometric group(VGG-19)

```
    print()
    print("Epoch" + str(epoch) + ":")
    print("Training Accuracy: " + str(training_accuracy) + "    Validation Accuracy: " + str(valid_accuracy))
    print("Training Loss: " + str(training_loss) + "    Validation Loss: " + str(valid_loss))
    print()
```

```
Epoch0:
Training Accuracy: 97.45385105028645    Validation Accuracy: 98.16200390370852
Training Loss: 0.06170810547716998    Validation Loss: 0.05126743076412596
```

Figure 4.2: VGG-19

Residual network(ResNet-50)

```
    print()
    print("Epoch" + str(epoch) + ":")
    print("Training Accuracy: " + str(training_accuracy) + "    Validation Accuracy: " + str(valid_accuracy))
    print("Training Loss: " + str(training_loss) + "    Validation Loss: " + str(valid_loss))
    print()
```

```
Epoch0:
Training Accuracy: 99.83856947721345    Validation Accuracy: 58.620689655172406
Training Loss: 0.005541473036821867    Validation Loss: 4.51361702637273
```

Figure 4.3: Residual network(ResNet-50)

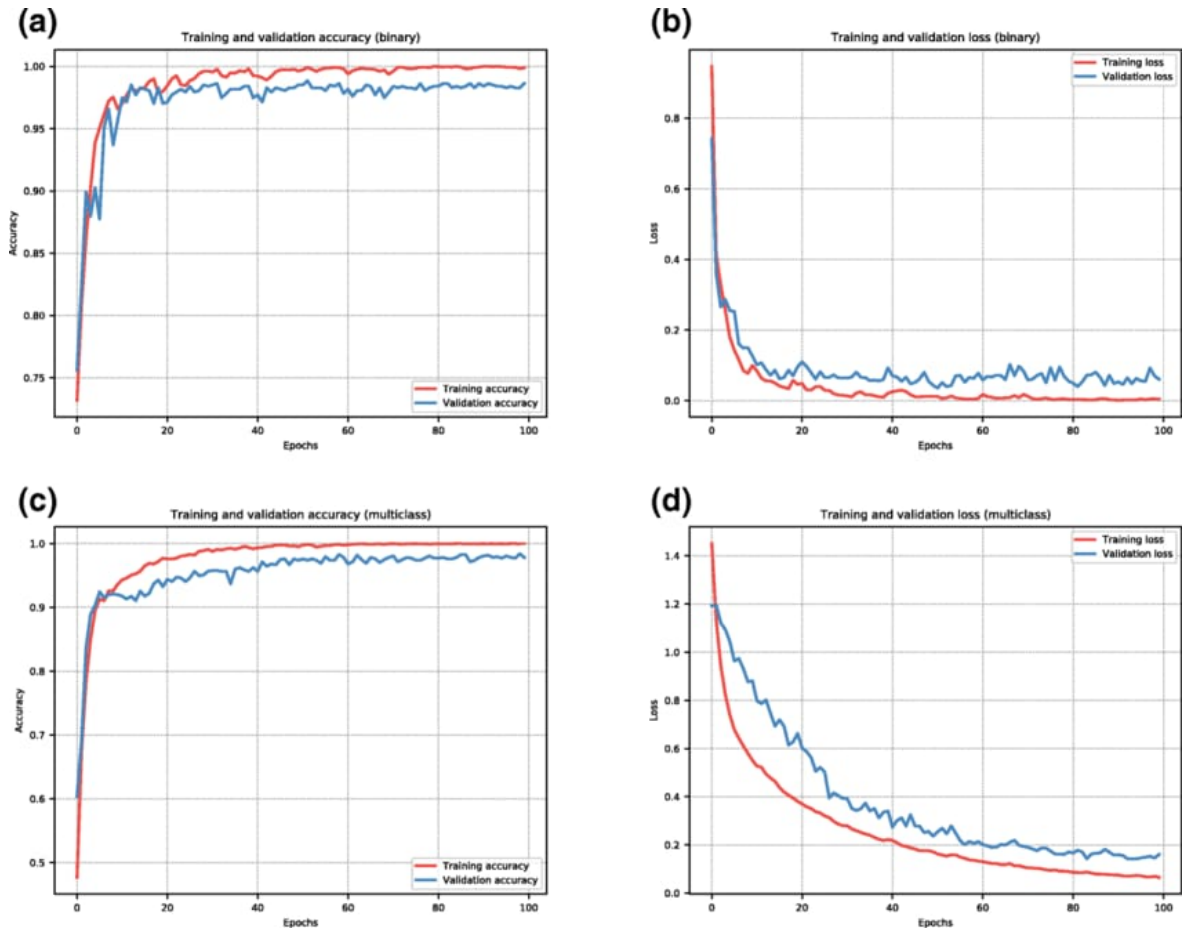

Figure 4.4: plots of CNN

REFERENCES

[1] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Commun.,* vol. 6, pp. 311-335, Mar. 1998.

[2] I. E. Teletar, "Capacity of multi-antenna Gaussian channels," *Europ. Trans. Telecom-mun.,* vol. 10, pp. 585-595, Nov./Dec. 1999.

[3] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: performance criterion and code construction," *IEEE Trans. Inform. Theory,* vol. 44, no. 2, pp. 744-765, Mar. 1998.

[4] M. R. Bell, J.-C. Guey, M. P. Fitz, and W. Kou, "Signal design for transmitter diversity wireless communication systems over Rayleigh fading channels," *IEEE Trans. Commun.,* vol. 47, Apr 1999.

[5] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Select. Areas Commun.,* vol. 16, pp. 1451-1458, Oct. 1998.*IEEE Trans. Commun.,* vol. 47, Apr 1999.

[6] B. A. Sethuraman, B. S. Rajan, and V. Shashidhar, "Full-diversity, high-rate space-time block codes from division algebras," *IEEE Trans. Inform. Theory,* vol. 49, no. 10, pp. 2596-2616, Oct. 2003.

[7] C.-C. Cheng and C.-C. Lu, "Space-time code design for CPFSK modulation over frequency-nonselective fading channels," *IEEE Trans. Commun.,*vol. 53, no. 9, pp.1477-1489,Sep. 2005.

[8] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2009)

[9] Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. arXiv preprint arXiv:1704.06904 (2017)

[10] Sanghyun, W., Soonmin, H., So, K.I.: Stairnet: Top-down semantic aggregation for accurate one shot detection. In: Proc. of Winter Conference on Applications of Computer Vision (WACV). (2018)

[11] He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Proc. of European Conf. on Computer Vision (ECCV). (2016)

[12] Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: Proc. of European Conf. on Computer Vision (ECCV). (2016)

[13] Bell, S., Lawrence Zitnick, C., Bala, K., Girshick, R.: Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2016)

[14] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Proc. of European Conf. on Computer Vision (ECCV). (2016)

[15] Chen, X., Gupta, A.: An implementation of faster rcnn with study for region sampling. arXiv preprint arXiv:1702.02138 (2017)

[16] Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. arXiv preprint arXiv:1608.06993 (2016)

[17] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proc. of Computer Vision and Pattern Recognition (CVPR). (2016)