CHAPTER I
INTRODUCTION

1.1. Background

The rapid advancement of artificial intelligence has brought significant changes to the way computers understand visual data. Deep learning is defined as a machine learning approach that enables computers to learn hierarchical data representations, where each layer constructs increasingly abstract representations from the preceding layer [1]. Furthermore, this deep learning approach has revolutionized various fields, including image recognition, speech recognition, and natural language processing, through its ability to automatically discover complex structures in high-dimensional data without requiring manual feature engineering [2].

One of the most influential deep learning architectures in the field of computer vision is the Convolutional Neural Network (CNN), which empirically demonstrated that CNNs trained on large-scale datasets are capable of achieving image classification performance far superior to conventional methods by significantly reducing error rates in the ImageNet Large Scale Visual Recognition Challenge [3]. This achievement marked the beginning of the exploration of increasingly deeper and more complex CNN architectures until the introduction of the Residual Network (ResNet), which addressed the gradient degradation problem in very deep networks through a shortcut connection mechanism. This innovation enabled the training of networks consisting of hundreds of layers and significantly improved feature representation capabilities across various computer vision tasks [4].

As CNN architectures became deeper, alternative approaches began to be explored to improve feature representation capabilities without continuously increasing the overall network depth. One approach that proved effective was the attention mechanism. This development led to the introduction of attention modules capable of explicitly modeling cross-channel dependencies through a squeeze operation using global average pooling and an excitation operation using two fully connected layers. These modules were shown to improve accuracy across various CNN architectures while introducing only a relatively small number of additional parameters. These findings demonstrated that modeling cross-channel relationships

alone provides substantial benefits for CNNs, which originally lacked such feature selection mechanisms [5]. Subsequently, attention mechanisms were further extended by sequentially combining channel attention and spatial attention, enabling networks to focus on both what is important and where important information is located within feature maps. It was concluded that combining multiple complementary attention mechanisms yields better performance than relying on a single attention mechanism alone [6]. Research on attention modules continued and revealed that the channel dimensionality reduction commonly employed in Squeeze-and-Excitation Networks introduces undesirable side effects. Local cross-channel interactions without dimensionality reduction, implemented through one-dimensional convolution, were found to achieve superior performance while maintaining lower computational complexity [7]. The evolutionary progression from SE-Net to CBAM and subsequently to ECA-Net collectively supports a solid conclusion: attention mechanisms are effective components for enhancing CNN feature representation capabilities, the combination of multiple complementary attention dimensions leads to improved performance, and the design of channel attention, including whether dimensionality reduction is applied, has a measurable and significant impact on the quality of the resulting feature representations.

In many previous attention modules, the separation of attention processing between the channel and spatial dimensions was generally adopted to maintain computational efficiency and limit parameter growth rather than because it represented the most optimal approach for feature representation [6]. In line with this observation, the study entitled “SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks” addressed this issue through a different approach. SimAM computes attention weights based on an energy function inspired by the spatial suppression mechanism of the biological nervous system and derives a direct mathematical solution (closed-form solution), allowing it to be implemented without introducing additional parameters into the network while generating three-dimensional attention weights in a simple and efficient manner [8].

Nevertheless, the mathematical formulation of SimAM contains a fundamental limitation that can be analytically identified as a consequence of its zero-parameter design. This limitation manifests itself in two distinct dimensions. First, the importance of each neuron is determined solely by the statistical distribution within an individual

channel without modeling cross-channel dependencies. Second, all neurons within a channel are averaged without considering their relative spatial positions, preventing the module from capturing spatial information at multiple scales. The first limitation is supported by findings from SE-Net, which demonstrated that cross-channel relationships captured only implicitly and locally by standard convolution operations constitute a significant weakness. Explicit modeling of cross-channel dependencies using global information was shown to substantially improve the representational power of the network, indicating that neglecting cross-channel relationships, as occurs in SimAM's energy function, imposes a meaningful limitation on the quality of feature representations produced [5]. The second limitation is evidenced by CBAM, which validated that explicitly modeling spatial context through a spatial attention module result in more accurate localization representations, particularly for tasks requiring high positional sensitivity. If a locally broad receptive field has been shown to be important, then the SimAM approach, which relies on a single global scalar average of all neurons without considering any relative positional relationships among them, is clearly incapable of capturing the spatial information required to determine attention focus [6].

To address limitations related to cross-channel interaction and multi-scale representation, the study entitled "Efficient Multi-Scale Attention Module with Cross-Spatial Learning" proposed the EMA module. EMA was designed to preserve cross-channel information without performing dimensionality reduction through a feature grouping mechanism. It integrates parallel subnetworks employing 1×1 and 3×3 convolution kernels to simultaneously capture short-range and long-range information, while applying a cross-spatial learning method to ensure that information from both branches is optimally fused through matrix dot-product operations rather than simple averaging, which may discard useful information [9]. The study empirically demonstrated that EMA outperformed several previous attention modules in image classification and object detection tasks while maintaining acceptable computational complexity.

The combination of multiple attention mechanisms has been extensively investigated in previous studies. For example, the study entitled *"Fine-grained Image Classification Method Based on Hybrid Attention Module"* combined channel attention and spatial attention within a hybrid attention module (MA) and further

incorporated an attention erasure module (EA) based on ResNet-50. Experimental results showed improvements in classification accuracy when all attention modules were integrated, indicating that combining multiple complementary attention mechanisms contributes positively to feature representation capabilities in CNN-based image classification tasks [10]. The potential of combining SimAM and EMA has also been demonstrated in the context of object detection. Specifically, the study entitled *"YOLOv8 Architectural Scene Section Recognition Method Based on SimAM-EMA Hybrid Attention Mechanism"* successfully integrated SimAM and EMA into the YOLOv8 architecture and demonstrated that their combination achieved better performance than either module used independently [11]. However, that study focused on object detection within a highly specific application domain and therefore did not provide a comprehensive evaluation of the effectiveness of the SimAM-EMA hybrid attention module for image classification tasks using a more general-purpose dataset and a standardized backbone architecture.

The evaluation of large-scale image classification systems requires appropriate datasets and standardized evaluation protocols. Top-1 accuracy and Top-5 accuracy have been established as the primary benchmarks for measuring classification performance, where Top-1 accuracy represents the proportion of correctly classified samples based on the highest-probability prediction, and Top-5 accuracy measures the proportion of samples for which the correct label appears among the top five predicted classes [12]. In addition to accuracy metrics, computational efficiency constitutes an inseparable dimension of comprehensive model evaluation. Fair comparisons require reporting the number of parameters and GFLOPs alongside accuracy metrics, while inference latency on actual hardware is not always proportional to GFLOPs and therefore must be measured empirically [13].

Based on this evaluation framework, this study additionally reports the number of parameters, GFLOPs, and training time as computational efficiency metrics accompanying Top-1 and Top-5 accuracy. The dataset used in this research is Tiny ImageNet, a dataset with a resolution of 64×64 pixels consisting of 100,000 training images and 10,000 validation images distributed across 200 classes, with 500 training images per class. The motivation for employing machine learning-based image classification on this dataset arises from the inability of rule-based methods and manual feature engineering approaches to handle high intra-class variation, visual

ambiguity among semantically similar classes, and the complexity of pixel distributions across 200 diverse categories. Consequently, deep learning, with its ability to automatically learn hierarchical representations, has become the only approach empirically proven to scale effectively for classification problems of this nature. Therefore, Tiny ImageNet serves as an appropriate benchmark for evaluating the feature representation capabilities of classification models under computationally affordable conditions [14].

In this study, the hybrid model combines a pretrained ResNet-50 backbone with two randomly initialized attention modules, SimAM and EMA. Such a configuration requires a specialized training strategy because the backbone and attention modules possess fundamentally different initial representational conditions. According to the study "How Transferable Are Features in Deep Neural Networks?" [15], features learned through CNN pretraining are transferable, indicating that the backbone does not require aggressive parameter updates. Therefore, a differential learning rate strategy is adopted following the principles of discriminative fine-tuning [16]. Under this strategy, the pretrained backbone within the Hybrid model is optimized using a smaller learning rate than the randomly initialized EMA module and fully connected layers, thereby preventing optimization conflicts among components with different initial representational states. In addition, a gradual freeze strategy is applied to EMA during the first three training epochs, allowing the backbone to adapt to the dataset distribution before all parameters are jointly optimized. These two mechanisms differential learning rates and gradual freezing of EMA are not implemented as arbitrary additional treatments but rather constitute an integral part of the research object itself, namely, the investigation of how the SimAM-EMA hybrid module can be effectively integrated and optimally trained within a pretrained backbone architecture.

Based on the discussion above, a clear research gap can be identified. The study *"SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks"* contains a fundamental limitation arising from its zero-parameter design, manifested in two distinct dimensions. Theoretically, these limitations can be addressed by the mechanisms proposed in *"Efficient Multi-Scale Attention Module with Cross-Spatial Learning."* Therefore, this research analyzes a SimAM-EMA hybrid attention module that integrates both modules sequentially into a ResNet-50

backbone and evaluates its effectiveness on the Tiny ImageNet dataset. The objective is to develop a model capable of surpassing the performance of each individual module through complementary mechanisms that collectively address the identified limitations.

1.2. Research Problem

Based on the background presented above, this study formulates several research questions focusing on the integration mechanism as well as the performance and efficiency evaluation of the SimAM–EMA hybrid model on the ResNet-50 backbone.

1. How can the SimAM and EMA attention modules be integrated as a hybrid mechanism into the ResNet-50 backbone for image classification on the Tiny ImageNet dataset?
2. How does the Top-1 and Top-5 accuracy performance profile of the SimAM-EMA hybrid model on the ResNet-50 backbone compare with the ablation reference results consisting of the baseline ResNet-50, ResNet-50 with SimAM only, and ResNet-50 with *EMA-Only* on the Tiny ImageNet dataset?
3. How efficient is the SimAM-EMA hybrid model in terms of the number of parameters, GFLOPs, training time, and inference latency compared with the three ablation reference variants on the Tiny ImageNet dataset?

1.3. Research Objectives

In line with the research questions, this study aims to design, implement, and analyze a ResNet-50-based image classification model incorporating a SimAM–EMA hybrid attention module in order to evaluate the resulting performance improvements and computational efficiency.

1. To design and implement the hybrid integration of the SimAM and EMA attention modules into the ResNet-50 backbone for image classification tasks on the Tiny ImageNet dataset.
2. To measure and analyze the Top-1 and Top-5 accuracy performance of the SimAM-EMA hybrid model on the ResNet-50 backbone in comparison with the baseline ResNet-50, ResNet-50 with SimAM only, and ResNet-50 with *EMA-Only* on the Tiny ImageNet dataset.

3. To analyze the computational efficiency of the SimAM-EMA hybrid model in terms of the number of parameters, training time, inference latency, and GFLOPs compared with the three ablation reference variants on the Tiny ImageNet dataset.

1.4. Research Benefits

This study is expected to provide significant contributions to both theoretical and practical domains in the development of attention modules for Convolutional Neural Network (CNN) architectures with a ResNet-50 backbone, particularly through the hybrid integration of SimAM and EMA.

1. As a theoretical contribution, this study provides empirical evidence regarding the hybrid integration of the SimAM and EMA attention modules in addressing a fundamental limitation arising from SimAM's zero-parameter design, which manifests in two distinct dimensions. First, the importance of each neuron is determined solely based on the statistical distribution within a single channel without modeling cross-channel dependencies. Second, all neurons within a channel are averaged without considering their relative spatial positions, making it incapable of capturing spatial information at multiple scales. Furthermore, this study presents a structured ablation analysis of the contributions of each model component, namely the SimAM-EMA Hybrid model, SimAM, EMA, and the baseline ResNet-50, in image classification tasks. Therefore, the findings of this research are expected to enrich the understanding of complementary attention mechanisms within the ResNet-50 CNN architecture and serve as a reference for future studies involving the combination of attention modules on similar backbone architectures.
2. As a practical contribution, this study produces a ResNet-50-based image classification model equipped with a SimAM-EMA hybrid attention module that can serve as a foundation for developing visual recognition systems requiring a balance between classification accuracy and computational efficiency. In addition, this research provides practical implementation guidelines for integrating two architecturally different attention modules within a single ResNet-50 CNN backbone, enabling the proposed approach to be replicated and further developed on different datasets and backbone architectures.

1.5. Research Limitations

This study is designed as an ablation study aimed at evaluating the contributions of the SimAM and EMA modules to the overall ResNet-50 architecture, including the associated training strategies. Within this framework, the Hybrid model serves as the primary subject of investigation, while the Baseline, SimAM-Only, and EMA-Only models function as ablation control variants, each isolating a specific condition: the inherent capability of ResNet-50 without attention modules, the independent contribution of SimAM, and the independent contribution of EMA. Accordingly, several limitations should be considered when interpreting the findings of this study.

1. The backbone architecture is limited to ResNet-50

   This study exclusively employs ResNet-50 as the backbone architecture in the conducted ablation experiments. Although its bottleneck block structure facilitates the modular integration of attention mechanisms, the resulting findings cannot be assumed to generalize to other architectures. Backbones such as MobileNetV2, which utilizes depthwise separable convolutions, DenseNet, which relies on dense connectivity, and Transformer-based architectures such as Vision Transformer (ViT), which inherently incorporate attention mechanisms, possess fundamentally different feature propagation characteristics. Consequently, their responses to the addition of SimAM and EMA may exhibit substantially different behaviors.

2. The dataset is limited to Tiny ImageNet-200 with a resolution of 64×64 pixels

   Tiny ImageNet, consisting of 200 classes and a resolution of 64×64 pixels, was selected due to its relatively affordable computational requirements. However, this choice introduces notable limitations. The relatively low image resolution results in feature maps with limited spatial dimensions at each stage of the network, thereby restricting the ability of attention modules to fully exploit spatial contextual information. Consequently, the findings of this ablation study may not necessarily remain valid when replicated on higher-resolution datasets such as ImageNet-1K or on domain-specific datasets with substantially different visual distributions.

3. Attention integration positions are fixed

   SimAM and EMA are evaluated only within a single integration configuration selected based on theoretical justification rather than empirical architecture search.

Alternative configurations that may also be theoretically valid are not explored. Therefore, there is no guarantee that the selected configuration represents the optimal design. Any claim regarding the superiority of the chosen configuration can only be supported theoretically rather than through comparative empirical validation.

4. Training is conducted on a single GPU through Google Colaboratory

   The experiments in this study are conducted using a single GPU provided through Google Colaboratory. As a result, the measured training time and inference latency are influenced by the characteristics of the specific hardware and software environment used during experimentation. Consequently, the reported efficiency metrics may differ when implemented on alternative hardware platforms or distributed computing environments.

5. The evaluation is limited to image classification tasks

   This ablation study evaluates the contributions of SimAM and EMA exclusively in the context of multi-class image classification using Top-1 and Top-5 accuracy metrics. In practical applications, backbones equipped with attention modules are frequently utilized as feature extractors for other computer vision tasks, such as object detection and semantic segmentation, where feature representation quality is assessed through different criteria, including spatial localization precision and multi-scale representation capability. Whether SimAM and EMA provide consistent benefits in such tasks cannot be concluded from this study and requires separate investigation.

6. The study is designed as an ablation study with predefined configurations

   This research focuses on an ablation study comparing the contributions of SimAM, EMA, and their combination within a controlled experimental framework. All models employ the same dataset, preprocessing procedures, optimizer, learning rate scheduler, number of training epochs, and training hyperparameters. However, they differ in terms of attention module integration strategies, freezing mechanisms, and differential learning rate configurations, which constitute integral components of each experimental scenario. Therefore, the obtained results reflect the relative effectiveness of the evaluated configurations rather than the best possible performance that could be achieved if each model were independently optimized through dedicated hyperparameter search procedures.