CHAPTER I
INTRODUCTION

This chapter provides a concise overview of the research in six sections. It commences by presenting the background and identifying the problem and research objectives. The chapter then proceeds to cover the hypotheses, research methodology, and problem limitations. A comprehensive explanation will be provided in the subsequent chapter.

1.1 Background

In computer vision, recognizing objects precisely is a challenging task. Fine-grained visual classification (FGVC) is a specialized approach designed to address this problem, focusing on distinguishing objects from nearly identical subcategories. Examples include distinguishing between bird species that appear very similar or identifying similar vehicle models. FGVC attempts to mimic the human visual ability to recognize subtle differences in computer-based systems. This technique is particularly important in situations where subtle differences between categories have a large impact, such as in biodiversity conservation, medical diagnosis, and smart retail. For example, FGVC can help monitor endangered species [1], [2], detect diseases through medical image analysis [3], [4], [5], or group products based on small variations in design and color [6], [7].

However, FGVC presents more complex challenges than coarse-grained classification due to the need to distinguish subtle differences within highly similar categories. The key distinction between coarse-grained image classification and FGVC lies in the level of detail and complexity. While coarse-grained classification sorts images into broad categories based on common features, FGVC delves deeper by identifying specific classes that differ in small subtle ways. FGVC is considerably more complex due to several factors: (i) substantial intra-class variation, where objects within the same class can appear in different poses and viewpoints; (ii) subtle inter-class differences, where minor details such as the color of a bird's head determine its class; and (iii) challenges in acquiring training data, as labeling fine-grained classes requires specialized expertise and significant annotation time, particularly for classes with limited data.

In FGVC, there are three main categories of strategies applied: part-based meth-

ods, feature encoding methods, and attention-based methods. Part-based methods focus on detecting and classifying discriminative areas within images. This approach emphasizes isolating and marking important parts related to specific objects. On the other hand, feature encoding methods concentrate on extracting high-level features to enhance object recognition accuracy. Meanwhile, attention-based methods leverage attention mechanisms to assess the importance of specific parts of objects, enabling the model to focus more on key details in images.

With advancements in FGVC research utilizing self-attention, researchers have begun to adapt transformers [8] for these tasks. Initially designed for text processing, transformers have become highly relevant for image recognition tasks. The Vision Transformer (ViT), introduced by Dosovitskiy et al. [9], represents a significant development in this architecture for image classification and object recognition tasks [10]. ViT segments images into patches that are then transformed into tokens, which are utilized in the multi-head self-attention (MHSA) mechanism during the training process [11]. However, there are two main challenges to address when using the ViT model for FGVC. First, in complex datasets or images with cluttered backgrounds, the ViT model may not efficiently capture attention on critical regions when processing all token patches simultaneously. Second, ViT has limitations in expanding its receptive field, which can result in the loss of crucial local information.

Based on these considerations, recent research has refined ViT architecture for FGVC tasks by improving the balance between global and local information. TransFG [12] enhances ViT by using attention weights to identify key patch tokens but struggles with small resolution images and complex datasets. AFTrans [13] addresses this by introducing a Siamese-based selective attention module but may lead to unreliable attention maps. Xu et al. [14] proposed Internal Ensemble Learning Transformer (IELT), which utilizes attention heads as weak learners to enhance cross-layer feature quality. IELT operates by strengthening internal representations through multiple attention heads functioning as an internal ensemble, allowing the model to be more effective at recognizing subtle variations between categories. However, despite improving the model's ability to process image details, IELT still faces redundancy issues, where irrelevant information reduces the model's efficiency. On the other hand, Ye et al. [15] introduced Batch-based Dynamic Mask (BDMM), an algorithm that dynamically adjusts the masking ratio on input patches to reduce information redundancy. BDMM works by filtering out image areas that do not contribute significantly to classification, ensuring that only relevant information is processed by the model. This approach improves both the efficiency and

accuracy of the model in fine-grained visual classification tasks by focusing attention on important features while ignoring irrelevant data, ultimately leading to better performance in recognizing subtle differences between objects.

Building on the proposed approaches, this thesis combines IELT and BDMM as a new approach to improve the performance of fine-grained visual classification (FGVC). These methods were chosen based on prior research that highlighted their ability to enhance the model's capacity for handling highly detailed visual tasks. Nonetheless, challenges related to information redundancy continue to affect model efficiency. To address this, the thesis incorporates BDMM, which dynamically adjusts masking ratios to minimize redundancy in input patches, allowing the model to focus on the most important image features. The primary goal is to enhance accuracy and precision in detecting subtle object differences, with the study expected to make a contribution to advancing FGVC.

1.2 Problem Identification

Based on the given background, it is imperative that the flaws in Dosovitskiy et al.'s introduction of ViT be addressed because of its use in the present FGVC. ViT uses the self-attention mechanism to analyze and categorize pictures, which yields substantial results and advances in image recognition architecture. However, there are still several fundamental issues that need to be resolved in order to enhance ViT's performance in FGVC tasks. A primary source of difficulty is the Multi-Head Self-Attention (MHSA) mechanism in ViT, which leans toward emphasizing global feature knowledge over local characteristics that are critical to FGVC. When ViT analyzes all tokens at once, it often fails to give enough attention to the crucial features required to distinguish between objects with minute variations. As a result, it is often unable to identify subtle but crucial differences, such color or shape, between similar objects. For example, it can't tell the difference between minute changes in a bird's feathers or beak. This restriction reduces ViT's ability to accurately categorize items that have extremely minor variations [12], [13], [14].

Furthermore, there exists an unevenness in the contributions of the several heads in the MHSA. Predictions made by certain heads may be inconsistent and erroneous because they are less adept at extracting significant aspects, while others may place an excessive amount of emphasis on unimportant details. This disparity raises the computational load without correspondingly improving efficiency, which in turn weakens the model's capacity to differentiate minute features [14]. In addition, the IELT model, which was developed to address some of these challenges, also

has weaknesses that need to be improved. Although IELT improves the model's ability to recognize fine variations between categories through internal ensembles, redundancy issues can still occur when generating cross-layer features, which are inputs to layer $L+1$. This not only reduces the efficiency of the model, but can also lead to inappropriate attention to important features resulting in performance degradation, especially in FGVC tasks that require precision in distinguishing small details.

This thesis proposes modifications to the multi-head voting (MHV) module of the IELT to address redundancy and noise issues by adjusting the convolutional kernel and applying a masking procedure. These modifications aim to improve the responsiveness of the model to subtle local variations and eliminate irrelevant data. According to [16], the kernel plays an important role in shaping the feature representation obtained by the model. In addition, Xu et al. [14] tested several types of kernels to optimize the performance of the IELT model. Therefore, in this thesis, additional types of kernels will be tested to ensure a more optimal convolution matrix/kernel captures more relevant discriminative features. In addition, the masking procedure is expected to help the model focus attention on significant local features by filtering out unnecessary information.

1.3 Objectives

The main goal of this research, as stated in the problem description, is to assess and analyze the ViT's performance, especially that of IELT model in the context of FGVC tasks, especially after adjustments have been made to improve the model's capacity to handle fine-grained features. The goal of the extremely difficult FGVC classification challenge is for the model to be able to discriminate between items with minute variations, such almost identical bird species or comparable flower varieties. In these kinds of jobs, the model must be able to capture and analyze the minute characteristics that are the main means of distinguishing across classes. To enhance the IELT's performance in FGVC, this study suggests two key changes. First, adjustments to the kernel in the MHV module are expected to enhance the model's ability to capture fine details that are often overlooked. These modifications aim to improve the model's efficiency and accuracy in FGVC, ultimately leading to more accurate object classification. Second, a masking technique is applied to the MHV module to filter out redundant information, ensuring that only the most critical features are considered in the voting process.

1.4 Hypotheses

Based on the problem statement, the hypothesis of this study is that modifying the MHV module in IELT by adding a masking process and modifying the kernel will help the model to learn and select important features in FGVC tasks more accurately. These modifications are expected to make the model focus more on key local features, such as eyes, beaks, feathers in birds, or certain fur patterns and facial structures in cat and dog datasets, thus improving the model's ability to distinguish objects with subtle differences as well as improving metric values, such as accuracy, precision, recall, and f1 score. As a result, the model is expected to be more adaptable to different types of objects in FGVC, whether distinguishing between bird species or cat and dog breeds, and more resilient to image noise, ensuring that it remains focused on the important features required.

1.5 Research Methodology

In this thesis, a fundamental study and experimental approach is used, which includes several important stages. First, a literature review is conducted to understand the basic concepts related to fine-grained visual classification, Vision Transformer, and other relevant techniques. This step aims to build a strong theoretical foundation to determine the model to be used for further research. Next, a suitable ViT model is selected and configured, including hyperparameter settings and training configuration, to ensure optimization during the training process. The next stage is the development of modifications and enhancements to the ViT model as well as the IELT model to improve its ability to understand local features in images. These enhancements are expected to improve the performance of the models in more detailed object recognition tasks. The modified IELT model will be trained using several datasets that show significant intra-class and inter-class variations, using appropriate performance metrics to measure the effectiveness of the proposed improvements. Finally, the obtained data will be analyzed and compared with previous methods to evaluate the extent to which the proposed method successfully addresses the problem of understanding local features in input images.

1.6 Problem Limitations

This research has several limitations that need to be considered in the effort to enhance the performance of ViT in FGVC tasks:

1. The research uses specific datasets for FGVC, such as CUB-200-2011 and Stanford Dogs. The results obtained may not be directly generalizable to other datasets with different characteristics.

2. The model training process only utilizes classification labels without additional annotations.

3. The foundational model used is ViT-B-16, which has been pre-trained on the ImageNet21K dataset.

4. During training, techniques such as random cropping, horizontal flipping, and color augmentation were applied. However, during testing, center cropping was used.

5. For learning rate scheduling, the cosine annealing method is used with initial learning rate values varying based on the dataset.