CHAPTER 1
INTRODUCTION

1.1 Background

The rapid advancement of computer vision technology has significantly impacted various aspects of human life, particularly in tasks such as image classification. As one of the fundamental tasks in this field, image classification has developed from Coarse-Grained Visual Classification (CGVC) [1–9], which can classify objects with coarse or obvious difference characteristics, such as birds, planes, and cars, to Fine-Grained Visual Classification (FGVC) [10–27], where classification tasks can distinguish objects with fine or subtle difference characteristics between sub-categories such as Field Sparrow and Grasshopper Sparrow bird, thus requiring the ability to distinguish objects based on small features that are difficult to identify using CGVC models or even by humans. Despite its advantages, FGVC presents three main challenges: (i) limited training data, since acquiring labeled data for fine-grained categories often requires domain expertise and substantial annotation time, thus restricting both the quantity and diversity of available data; (ii) high intra-class variation, where significant differences in object appearance caused by changes in position, viewpoint, or lighting condition within the same category, making intra-class consistency difficult to capture; and (iii) low inter-class variance, in which different categories exhibit highly similar visual features except for minor distinguishing parts, complicating accurate discrimination between classes.

Proposed to solve the first challenge is Weakly Supervised Learning (WSL) [28]. WSL uses high-level labels such as image classes instead of depending on thorough and localized object annotations such as bounding boxes or part labels to enable model training. Using this approach significantly reduces the dependability on labor-intensive manual labeling processes. It lowers the cost and time required for data collection, making it suitable for domains where expert annotation is challenging. However, WSL alone cannot adequately resolve the issues of high intra-class variation and low inter-class variance, as it lacks mechanisms to guide the model toward consistently focusing on subtle and discriminative object parts necessary for the FGVC task.

The Weakly Supervised Data Augmentation Network (WSDAN) [18] integrates WSL with attention-guided data augmentation to address the remaining challenges.

This approach employs two augmentation strategies. For intra-class variation, the model may produce wrong predictions if it only focuses on some parts of the object, especially when variations hinder those parts, so the attention-dropping function removes a salient region from the image, encouraging the model to extract discriminative features from other relevant parts. Attention cropping and resizing essential discriminative areas for inter-class similarity help the model learn minute but significant variations between classes. WSDAN has limits even if it offers a potential solution by merging augmentation with attention mechanisms. Specifically, it relies solely on the final prediction outcome without evaluating the causal relationship between the attention and prediction results. Consequently, the model lacks guidance on which regions are truly discriminative and which may lead to biased or incorrect predictions.

To overcome this limitation, Counterfactual Attention Learning (CAL) [23] was introduced. CAL enhances attention learning through the lens of counterfactual reasoning, a principle that has shown effectiveness in various machine learning applications, starting from image classification [29, 30], object detection [31], object recognition [32], reinforcement learning [33], Natural Language Processing (NLP) [34], Visual Question Answering (VQA) [35, 36], Vision-and-Language Navigation (VLN) [37], scene understanding [38], to recommender systems [39]. In CAL, attention quality is evaluated by comparing the prediction outcomes of two settings: real (factual) attention and artificially generated fake (counterfactual) attention. The difference between these two outcomes, called the causal effect, helps the model identify meaningful attention regions and mitigate bias during training. In the original research, CAL achieves the best accuracy by using a uniform distribution to generate random attention as a form of counterfactual. However, CAL does not explicitly specify or analyze the impact of the distribution type used, which may not necessarily be the most effective choice. The selection of counterfactual attention types could significantly influence model performance. Using random distributions as fake attention is an innovative approach to improve the model's robustness to training bias. Although fake attention is generated through a Global Average Pooling (GAP) process between the initial fake and the real attention, this approach remains challenging. The main challenge lies in the potential uncertainty introduced by the random distribution, which can affect the model's performance depending on how the attention is generated and integrated. To address this challenge, this study proposes an annealing-based mechanism called Annealed Counterfactual Attention (ACA), a stepwise strategy inspired by the material cooling process in metallurgy, where a system transitions from a highly variable state to a more stable one.

This approach has been effectively integrated into tasks ranging from reinforcement learning to neural architecture optimization, indicating the annealing mechanism's broad applicability and adaptability in complex learning scenarios, as shown in prior works [40–44]. In ACA, this principle is applied by progressively replacing fake attention with real attention during training. This allows the model to benefit from a random distribution to mitigate bias in the early phases and gradually concentrate on real attention to enhance accuracy and stability.

In addition, several things will be taken in this study to improve the performance of the CAL model. So, the contribution of this study can be summarised as follows:

1. Analyzing the effects of various random distribution types used to generate fake attention maps to identify the most effective type for counterfactual implementation.

2. Proposing Annealed Counterfactual Attention (ACA), a novel annealing-based mechanism that progressively transitions fake attention to real attention during training, improving both model robustness, accuracy, and certainty.

The rest of this study is organized as follows. Section II describes the basic concepts. Section III describes the proposed method. Section IV describes the experiments and results, followed by the conclusion of the study in Section V.

1.2 Problem Identification

In the CAL Method [23], four types of counterfactual attention are used to implement the counterfactual principle: random attention, uniform attention, reversed attention, and shuffle attention. Among these, experimental results show that the random attention counterfactual type produces the best accuracy compared to other types of counterfactual attention. However, its performance still leaves room for improvement, thus requiring further analysis of the counterfactual random attention distribution type. While the best distribution types can improve model accuracy, using a random distribution introduces uncertainty, as the final accuracy depends on the generated distribution. Therefore, a novel approach is needed to achieve a model that produces more stable and deterministic results.

1.3 Objective and Contributions

This study aims to improve the accuracy and robustness of FGVC by developing a novel attention mechanism called Annealed Counterfactual Attention (ACA).

ACA is designed to enhance the Counterfactual Attention Learning (CAL) framework by introducing an annealing-based strategy that gradually transitions from fake (counterfactual) attention to real (factual) attention during training. Inspired by the annealing principle, this mechanism enables the model to initially benefit from random concerns to reduce bias and progressively refine its attention to focus on truly discriminative regions, consequently improving stability and classification performance. This study analyzes several types of random distributions used to generate fake attention maps in order to maximize the performance of ACA. The aim is to find the most efficient distribution that functions as the initial counterfactual attention in the annealing process, optimizing the overall advantage of ACA in the training dynamics.

1.4 Scope of Work

The scope of work is described as follows:

1. The programming language used for model implementation is Python, utilizing PyTorch as the deep learning framework.
2. The proposed Annealed Counterfactual Attention (ACA) mechanism is implemented on top of the Counterfactual Attention Learning (CAL) framework.
3. Only the best-performing random distribution type is used as the initial counterfactual attention in ACA.
4. The datasets used for FGVC tasks are FGVC-Aircraft [45] and CUB-200-2011 [46].
5. The evaluation metrics used is top-1 accuracy.

1.5 Expected Results

The expected outcome of this study is to achieve good results in developing the Annealed Counterfactual Attention (ACA) mechanism, successfully identifying the most effective type of counterfactual random attention distribution, and demonstrating that the proposed ACA model can achieve higher accuracy, robustness, and overall performance compared to the original CAL model in FGVC tasks.

1.6 Research Methodology

In this study, studies and experiments based on Work Packages (WP) are used. The following are the WPs for this study:

- WP 1: Conduct a literature review on the topics of FGVC task, counterfactual, and annealing method.
- WP 2: Perform model selection and configuration settings such as hyperparameters and training settings.
- WP 3: Implement contributions to improve the model performance.
- WP 4: Evaluate the model performance on two benchmark datasets.
- WP 5: Analyze the improved model and compare with the other State-of-the-Art (SOTA) methods.