ABSTRACT

Image analysis and visual recognition play a vital role in various modern applications, including face recognition, autonomous vehicles, and object detection. Fine-grained visual classification (FGVC) is a significant challenge in visual classification, where the ability to distinguish fine details among visually very similar objects is a major constraint. Recently, Vision Transformer has emerged as one of the methods used for visual tasks. Although Vision Transformers (ViTs) have shown great potential in various visual tasks, there are still limitations in understanding and extracting local features from images, which are the main components in FGVC.

In this thesis, we modified the Internal Ensemble Learning Transformer method by modifying the Multi-Head Voting (MHV) module by integrating more appropriate kernels and applying masking techniques through the implementation of the Batch-based Dynamic Masking (BDMM) algorithm to improve the model's ability to understand and extract local features from input images, thereby improving accuracy and other metrics. We conduct several testing scenarios to find the best method. First, we explore various types of kernels contained in the MHV module in the IELT. In the second scenario, we explored the placement of masking techniques in the IELT method. Then, we combine the best results from both scenarios, which result in significant performance improvements in terms of accuracy, precision, and efficiency.

The results obtained show that each scenario consistently improves the accuracy of all tested datasets. Kernel sharpening shows higher accuracy results compared to other kernels. In addition, the placement of the masking method in the MHV module also provides an increase in accuracy. The best results are obtained by the combination of convolution kernel modification with kernel sharpening integrated with the BDMM algorithm, which provides an accuracy increase of 0.1% to 0.3% on all tested datasets when compared to other methods that have been trained using the same dataset.

Keywords: Fine-Grained Visual Classification, Vison Transformer, Internal Ensemble Learning Visual Transformer.