ABSTRACT

Research in image classification has attracted considerable interest, particularly in Fine-Grained Visual Classification (FGVC), specializing in the complex task of differentiating objects and subtle variations within species, such as those observed in different animal types. Various techniques have been developed to tackle these challenges, including feature coding, part-based approaches, and attention-based approaches. Within these approaches, the Vision Transformer (ViT) has shown remarkable success in image recognition tasks. The Internal Ensemble Learning Transformer (IELT) builds on ViT as its foundation, achieving impressive results. To enhance the effectiveness of IELT, we propose an innovative approach that focuses on refining its feature representation. This is achieved by incorporating a softmax activation function and a Radial Basis Function (RBF) layer to enhance the final prediction accuracy. Experimental findings demonstrate that our proposed method significantly boosts accuracy on fine-grained datasets, such as Oxford-IIIT Pet, Surpassing current cutting-edge methods.

Keywords: Fine-Grained Visual Classification, Vision Transformer, Radial Basis Function.