RA-FGIC: Retrieval Augmented fine grained Image Classification

This research explores the potential of Retrieval-Augmented Fine-Grained Image Classification (RA-FGIC) for overcoming limitations in traditional FGIC methods. Fine-grained image classification (FGIC) presents a significant challenge due to subtle visual variations within categories. This work proposes a novel RA-FGIC approach that leverages a CLIP encoder for semantic image representation and a Convolutional Neural Network (CNN) for feature extraction. During classification, the model retrieves similar images from a CLIP-encoded vector database for a given image. To capture the interaction between the target image and retrieved images, we propose a cross-product operation between the retrieved image embeddings and the last layer output of the pretrained CNN of target image. This enriched feature representation is then fed to a classification layer for fine-grained category prediction.

We evaluate our approach on benchmark datasets such as Stanford Cars and investigate the effectiveness of the cross-product strategy compared to alternative feature combination methods. We evaluate our approach on these datasets and achieve significant improvements in classification accuracy, surpassing a pre-trained EffecientNet CNN model by 11%. Our work demonstrates the effectiveness of RA-FGIC with CLIP and cross-product feature interaction for fine-grained image classification tasks.