Rag Enhanced Image Classification

This project investigates Retrieval-Augmented Generation (RAG) for image classification by integrating Vision Transformers (ViTs) and k-Nearest Neighbors (KNN) search to enhance classification accuracy. Using a pre-trained ViT-Tiny model, we extract CLS and patch embeddings from training images for CIFAR-10 classification.

We propose two primary approaches for sign language recognition: (1) CLS-Based Classification, where each image is represented by a single CLS token, and classification is performed using a KNN approach based on CLS embeddings from the training set, with the final class determined by majority voting; and (2) Patch-Based Classification, where each image is divided into 196 patches and KNN is applied at the patch level. Each patch is classified based on the majority vote among its k-nearest neighbors, followed by a second majority vote across all patches to determine the final classification.

Additionally, we extract softmax predictions from a fine-tuned ViT-Tiny classifier and incorporate them into a prediction-with-function method. This function computes the final classification by summing the contributions of the K nearest neighbors, where each neighbor's class label is weighted by its similarity to the test image. The similarity measure determines how close the test image embedding is to each training image embedding and this value is multiplied by the corresponding one-hot class vector. A static distance factor, representing a perfect match contribution, is also incorporated along with the predicted softmax vector from the fine-tuned Vision Transformer. This term ensures that the global prediction from the Vision Transformer plays a role in refining the decision, balancing both local instance-based classification and global feature-based recognition for improved accuracy.

Our experiments show that cosine similarity marginally outperforms Euclidean distance, increasing K improves accuracy up to an optimal threshold at K = 9–13, integrating softmax probabilities unexpectedly reduces accuracy due to misalignment between probability distributions and similarity-based classification and CLS version gives better results than the Patch-based version. Future work will explore distance weighting, alternative similarity metrics and additional techniques to enhance performance.