Semantic Image Retrieval

Introduction

Semantic image retrieval uses advanced algorithms and deep learning techniques to find images that match the user's query based on the understanding of the content and context within the images. The primary objective of my internship was to develop a solution that enables efficient image search based on images, moving beyond traditional methods that rely solely on metadata or manual labeling, which can be prone to errors.

Project Overview

The project was divided into two main phases:

1. **Image Captioning**: The first phase focused on generating accurate and detailed captions for images using state-of-the-art models.
2. **Image Retrieval**: The second phase involved retrieving images based on user queries by embedding both the captions and queries in a shared vector space and performing a similarity search.

Image Captioning

For the image captioning phase, we evaluated multiple state-of-the-art models, including Florence 2[1], LLaVA[2], and GPT-4o. In our experiments, we explored various approaches to caption generation, including direct captioning from the models and using carefully designed prompts to guide the captioning process. Through qualitative and quantitative assessments using metrics such as BLEU and ROUGE, we found that detailed captions, especially those generated using tailored prompts, provided more useful information for retrieval tasks. The final model selection was based on a balance between the detail provided in the captions and computational efficiency. Models that could generate high-quality, context-aware captions without excessive processing overhead were prioritized for the final implementation.

Image Retrieval Workflow

We experimented with various embedding models, including Alibaba's GTE Large, Intfloat's Multilingual E5 Large Instruct, and OpenAI's Text Embedding 3 Large. The best performers were selected based on their ranking in the MTEB (Massive Text Embedding Benchmark)[3] leaderboard. The MTEB provides a comprehensive evaluation across various embedding tasks, ensuring that our chosen models offered robust performance across different scenarios. The similarity search was primarily conducted using cosine similarity to match the user query with the stored image descriptions.

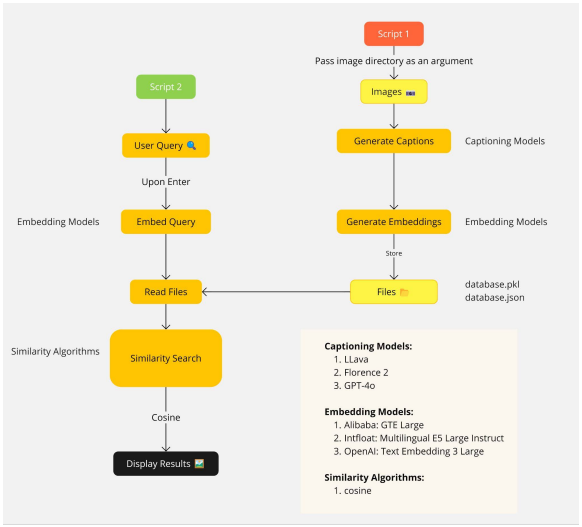

Figure 1: Image Retrieval Workflow.

Results

After extensive testing and evaluation, we successfully identified the best combination of captioning and embedding models for our image retrieval system. These combinations provided an optimal balance between accuracy, computational efficiency, and resource usage. The proof of concept (POC) was developed as a web-based application using Streamlit, demonstrating the image retrieval solution for DxO's software like PhotoLab.

References

[1] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, Lu Yuan, "Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks", arXiv preprint arXiv:2311.06242, 2023.

[2] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, Chunyuan Li, "LLaVA-OneVision: Easy Visual Task Transfer", arXiv preprint arXiv:2408.03326, 2024.

[3] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, Nils Reimers, "MTEB: Massive Text Embedding Benchmark", arXiv preprint arXiv:2210.07316, 2022.