CS3570 Multimedia Final Project Proposal

Image Retrieval

吳善凱 109020003 李承澔 109020014 蘇暐中 109020021

Introduction

Image Retrieval is a computer vision task that involves finding images similar to a provided query from a large database. This final project aims to employ the CLIP (Contrastive Language–Image Pre-training) model proposed by OpenAI, to accurately identify and retrieve the most relevant images based on textual query.

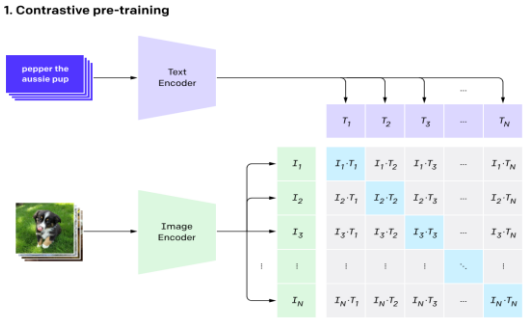
We will utilize the condensed version of iNaturalist Dataset released by TA, which contains 10000 images for training, 1000 for testing, and 10 categories.

CLIP (Contrastive Language–Image Pre-training) Model

The fundamental idea of CLIP is the alignment of images with their corresponding textual descriptions within a shared embedding space. CLIP contains 2 encoders: one for text, transforming descriptions into vector embeddings, and one for images, mapping them into the same vector space.

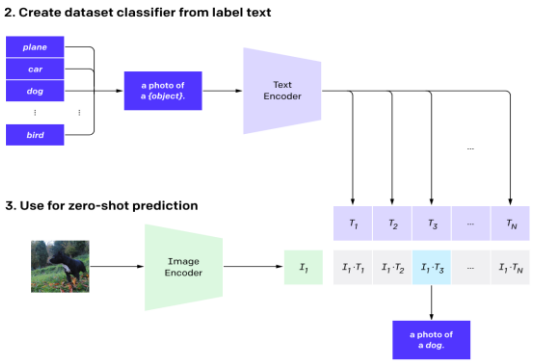
For a given batch of ***n*** corresponding images and texts, the encoders generate mappings to vectors

(T1, T2, T3, …, Tn) for text and ***(I1, I2, I3, …, In)*** for images.



For each paired transformation ***(Ti, Ii)***, we expect a higher similarity, represented by an inner product with a lower cosine angle between the vectors. To achieve this, CLIP utilizes a loss function predicted on cosine similarity.

Upon completion of training, CLIP projects all textual descriptions of categories and a query image into this shared space. The textual category which yields the highest inner product with the query image, is viewed as the most similar category.



System Design

There are three stages in the system: preprocessing, training, and inference.

The initial step involves preprocessing the data, which includes building a data loader and ensuring the proper connection between category labels and images.

For the training phase, we will use the CLIP LAION ViT-H/14 model, which is trained with 2B samples, as the pretrained model to fine-tune our system. This involves building a training script that leverages the pretrained model to learn from our specific dataset and accurately align textual descriptions with their corresponding images in the shared embedding space.

During the inference stage, the first step is to generate vector embeddings for all testing data. The system will project each image and its corresponding textual description into vector embeddings within the shared embedding space.

To retrieve the most relevant or top similar images based on a query, the system will generate a vector embedding for

the query and then find the most similar or top similar vector embeddings within the vector database. The output will be the original image(s) corresponding to the obtained vector, identifying the most relevant image(s) based on the provided image query.

Reference

laion/CLIP-ViT-H-14-laion2B-s32B-b79K

openai/CLIP

mlfoundations/open_clip

Hugging Face documentation of CLIP

CLIP