Image Retrieval via Multimedia Encoding

Sharvi Tomar
stomar2@illinois.edu

September 24, 2022

1 Introduction

This document reports on the Image Retrieval System that I developed using multimedia embeddings. Given the image, it is able to retrieve corresponding captions/text descriptions, and vice versa. The main idea is to represent each image and each caption using multimedia encoders, and find the captions (or images) most similar to a query image (or caption).

2 Approach

Adopted Model CLIP (Contrastive Language Image Pretraining)

The CLIP model from OpenAI is for connecting images and caption texts. This massive model was trained on 400M pairs of images and captions trained on the web. CLIP aims to find a mutual latent space for Images and Text prompts. This is done by taking pairs of Images with their matching Captions, running them through some encoders (one Text Encoder and one Image Encoder respectively) and encouraging their cosine similarity to be high. They also use the other pairs in the batch as negative examples, encouraging their cosine similarity to be low. This results in a feature space that has a deep connection between a certain Image representation, and its matching Text.

I adopted the zero-shot learning CLIP model from OpenAI for retrieving image-to-text and text-to-image retrieval.

3 Dataset

Dataset: MSCOCO 2014 version

Splits: "dataset_coco.json"

from: `https://cs.stanford.edu/people/karpathy/deepimagesent/caption_datasets.zip`

Data info:

Train samples: 113,287

Test samples: 5000

Validation samples: 5000

Captions: 5 per image

4 Performance Analysis

Image Retrieval is commonly evaluated with Average Precision (AP) or Recall@k. In this homework, I computed Recall@1, Recall@5 and Recall@10 as well as Median Rank and Mean Rank for both image-to-text (i2t) retrieval and text-to-image (t2i) retrieval.

Recall at k is the proportion of relevant items found in the top-k recommendations

$$\text{Recall@k} = \frac{(\text{\# of recommended items @k that are relevant})}{(\text{total \# of relevant items})}$$

	recall@1	recall@5	recall@10	median_rank	mean_rank
image-to-text	34.72	70.3	80.4	2.0	10.6716
text-to-image	22.54	49.52	61.856	6.0	36.97248

The image-to-text model has a very good recall for 5 and even better 10 items and hence would be good to have 5 or 10 items recommendation.

The text-to-image model has a very good recall for 10 items and hence would be good to have 10 items recommendation.