Course Name: Computer Vision

Weekly Report: 2

Group Name: XYZ

Submitted to faculty:

Mehul Raval

Date of Submission:

29 Feb 2025

Student Details

Roll No.	Name of the student	Name of the program
AU2240106	Meet Rathi	B.Tech in CSE
AU2240160	Harsh Panchal	B.Tech in CSE
AU2240153	Aditya Agarwal	B.Tech in CSE

Table of Contents.

WORK DONE THIS WEEK

What we understood:

CLIP and the Multimodal Models

We studied about the Contrastive Language-Image Pretraining(CLIP) and multimodal models, they provides critical foundation for understanding text and image relations for our project. The vision-language model CLIP receives training as a pretrained system, it pairs images with their matching text descriptions.

The system follows multimodal learning by getting information from simultaneous text and image inputs. Instead of classifying images into specific categories, it determines proper matches between text and images while separating incompatible pairs. CLIP operates through a text encoder that relies on transformer technology to process natural texts efficiently while its image encoder utilizes the vision transformer to extract meaningful image representations.

It works with the help of the text and image encoders, which converts the inputs into the high dimensional vectors. They are further mapped to the shared space. It then increases or decreases the similarity based on the matched and mismatched pairs. After the training process, it can get the images based on the queries.

Reference:
https://www.marqo.ai/course/introduction-to-clip-and-multimodal-models

https://www.geeksforgeeks.org/clip-contrastive-language-image-pretraining/

EDA around the flickr dataset

- The Flickr Image Dataset available on Kaggle served as our dataset foundation which included pictures alongside their associated captions.
- We did Exploratory Data Analysis (EDA) on dataset to review its structure together with the distribution of the captions and examination of its metadata components.
- We observed variations in dataset in terms of image content, caption length, and descriptive quality. It helped us to understand, how well does this dataset aligns with the CLIP code and it's objectives
- We played around with this dataset, to understand how will we process this dataset for the CLIP model, like for example testing how will the captions influence the retrieval performance

Reference:

https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset

playing around the CLIP code on flickr dataset

- We understood the CLIP model by implementing its code. We played around with different data prompts to implement it.
- We tried to test its ability to recognize the semantic relationship between the text and the image.
- Through the hands-on experience with the clip code, we understood how it actually works.

WORK TO BE DONE NEXT WEEK

1. PRS dataset with CLIP: We will train the model efficiently of PRS dataset using CLIP, then with this we will test its ability to relate the textual and the visual data.