Ahmedabad
University

Course Name: Computer Vision

Weekly Report: 1

Group Name: XYZ

Submitted to faculty:

Mehul Raval

Date of Submission:

22 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

Our Problem Statement:

We are required to find the images based on text description. For example, “a person wearing a red jacket, a black cap and brown shoes”, and the model will find someone matching the description. This can be done using the CLIP model, which understands both text and the images.

Now running such model on small, low-power device, such as Nvidia Jetson Orin AGX would be challenging, as these devices struggle with speed and accuracy.

Our project aims to optimize the CLIP model using a technique known as (Low-Ranking Adaptation), to make it work on the small device provided to us.

What we understood:

Attention Mechanism and Transformer Technique:

We explored on the attention mechanism and transformer technique, as they are foundation of our project. We focused on how does this model helps in efficient processing of multimodal data in the model like CLIP.

- Attention Mechanism
    - Allows the model to focus on the most important parts of the data from input
    - Self-attention helps to identify the in-data relationships, which can be useful for the text and vision tasks.
    - Multi-head attention helps improve learning by processing the different features in parallel

- Transformer Technique
    - It avoids the sequential dependency and processes the input elements simultaneously
    - To maintain the sequence of the information, they use positional encoding
    - The key components are self-attention layers, feed-forward networks (FFN), layer normalization, and residual connections for the stability and better learning of data.
- In CLIP:
    - Transformers will help for the encoding of text and images into a shared representation space.
    - Applying this model will be able to learn unseen categories based on the text and image alignment

Reference:
https://uvadlc-notebooks.readthedocs.io/en/latest/tutorial_notebooks/tutorial6/Transformers_and_MHAttention.html

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/

LoRA(Low-Rank Adaptation)

It is a technique, in which instead of updating the entire model, it fine tunes the large pre-trained model, i.e. in our case CLIP, efficiently by using only a small no of parameters.

It works such that, instead of modifying all the params, it adds the low-rank matrices to the specific parts of the model. So further these small layers keep on capturing the task-specific knowledge where on the other hand the original model is frozen. At the inference time, what the model do is, that it combines the original model weights along with the small LoRA layers to generate the predictions.

So LoRA is important for the CLIP based person retrieval system on a small edge device, where it will help fine-tune this model, without the need of excessive computation power or memory.

Reference:
https://huggingface.co/docs/peft/main/en/conceptual_guides/lora

We also took the overview of the jetson orin agx architecture

Nvidia Jetson AGX orin

It is a computing module which is designed for the edge AI applications which requires the high performance, with the low power consumption.

ports:

Micro-B - USB-2.0 - for the serial debug console

DisplayPort 1.4a - only way to get the display output out of u the developer kit

Two USB TYPE-A ports - USB3.2 - GEN2

RJ45 jack for the ethernet

Power jack - 9-20V

USB type-c jack - USB-3.2 GEN2 - power the jetson from the power brick

Micro SD card slot

Buttons:

Power button

Force recovery button

Reset button

Reference: https://www.youtube.com/watch?v=LUxyNyCl4ro

WORK TO BE DONE NEXT WEEK

Will do exploratory data analysis on the dataset, and will try to structurise the dataprompts, and will try to make a pilot model with CLIP.