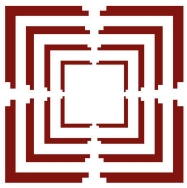


Ahmedabad
University

CSE641: Computer Vision: Modern Methods And Applications

Report-1

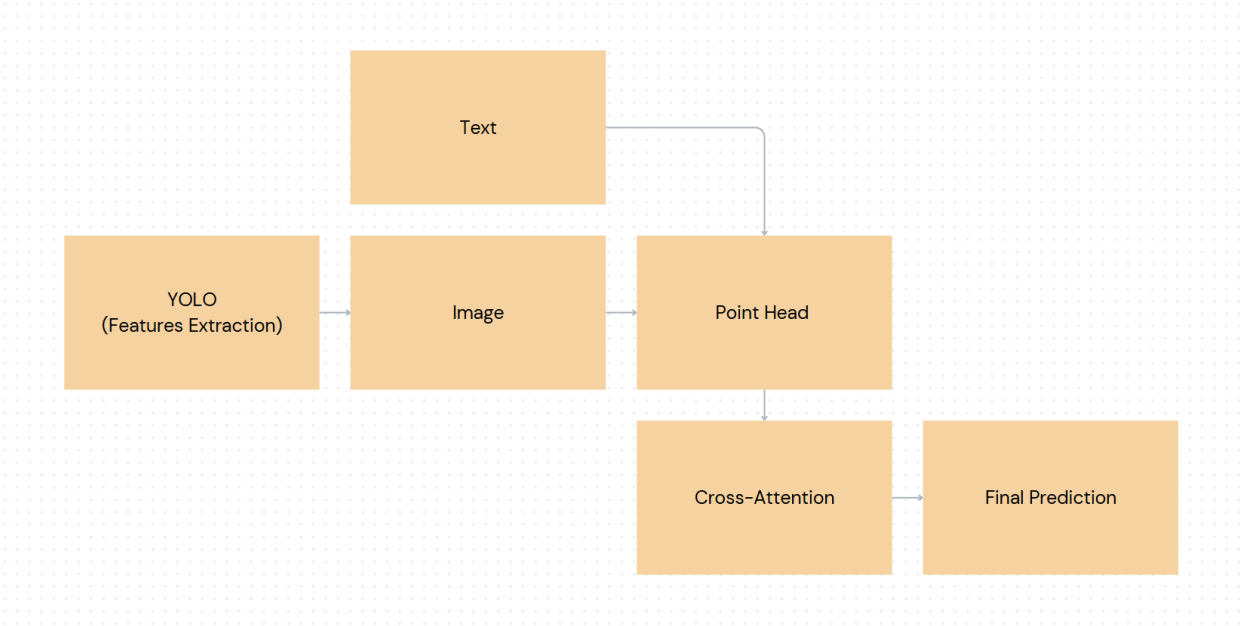
Group 1

Name	Enrollment No.
Dhyey Patel	AU2240054
Malav Modi	AU2240214
Prem Patel	AU2240010

Understanding Problem Statement:
The feature extraction portion uses YOLO to extract object information while CLIP provides multi-modal learning capabilities that enable similarity-based image-text alignment in solutions. The process first acquires features from YOLO before sending them to the image transformer block and following both text and image data to the point header before executing the cross-attention block. The pipeline determines similarities through the implementation of the cosine function. Researchers aim to explore various network models in the image block of their pipeline alongside finding the optimal model thus far.

Key Steps in the Pipeline:

- **Feature Extraction:**
  YOLO extracts object-centric features from images.
- **Text + Image Block**
  Processes text and Image.
- **Cross-Attention Block**
  A cross-attention module enhances both features while making them more extensive for better recognition.
- **Similarity Computation**
  The block determines image-text similarity through the application of cosine function.



Task completed:

- **Literature Review:**
  In order to increase model efficiency and generalization, recent developments in computer vision have investigated several pretraining techniques. Using a dataset of 400 million (text, picture) pairs, Radford et al. (2021) developed CLIP (Contrastive Language-picture Pretraining), which allows for zero-shot learning across a variety of tasks and learns visual representations from natural language supervision. Goldblum et al. (2023), on the other hand, carried out a comprehensive comparison of pretrained backbones, such as CNNs, Vision Transformers (ViTs), and self-supervised learning (SSL) models, across tasks like object identification, classification, and out-of-distribution generalization. Although SSL models become competitive when scaled suitably, their results show that supervised CNNs (e.g., ConvNeXt, SwinV2) still surpass transformers on the majority of tasks. Even though CLIP shows how effective vision-language learning can be for generalization, task-specific performance is still dominated by classic supervised and SSL-based methods, underscoring the significance of choosing the appropriate model design and training model for certain applications.

- **Understanding of the Code:**
  The code is to understand the relationship between images and text by implementing CLIP (Contrastive Language-Image Pre-training) . An overview of CLIP is given initially, and then dependencies like timm are installed. Important libraries like OpenCV, NumPy, Pandas, PyTorch, Albumentations (for image augmentations), and Matplotlib (for visualization) are imported into the notebook. In order to prepare inputs for training, it also incorporates data pre-processing processes, most likely requiring text tokenization and image transformations. Training or optimizing a CLIP model to understand image-text relationships appears to be the main objective.

Goals for next week:

- Analyse the pretrained models like EfficientNet, CoatNet and MobileNet and how we can implement it on our pipeline
- Run this models and compare the performance on Flicker dataset.
- Make this pipeline more efficient and prepare it for a Person Retrieval dataset.