Video indexing and retrieval using CLIP

Vinit Kujur (2021CSM1010)
Tauqeer Akhtar (2021CSM1021)

Introduction

In this project, we look at the task of video retrieval by text, where a query is described in the form of a natural-language sentence. The task is scientifically interesting and challenging as it requires establishing proper associations between visual and linguistic information presented in the temporal order. To get the relationship between the text and the image CLIP[1] (Contrastive Language–Image Pre-training) pretrained model is used. The CLIP model was trained on 400 million image-text pairs to efficiently learn visual concepts from natural language supervision (see Figure 2 for the model architecture). The significant frames or keyframes are extracted from the video using FFmpeg library to get overall content information about the video. For faster query inference time, we have stored the encoded features of keyframes from the video as meta for later use.

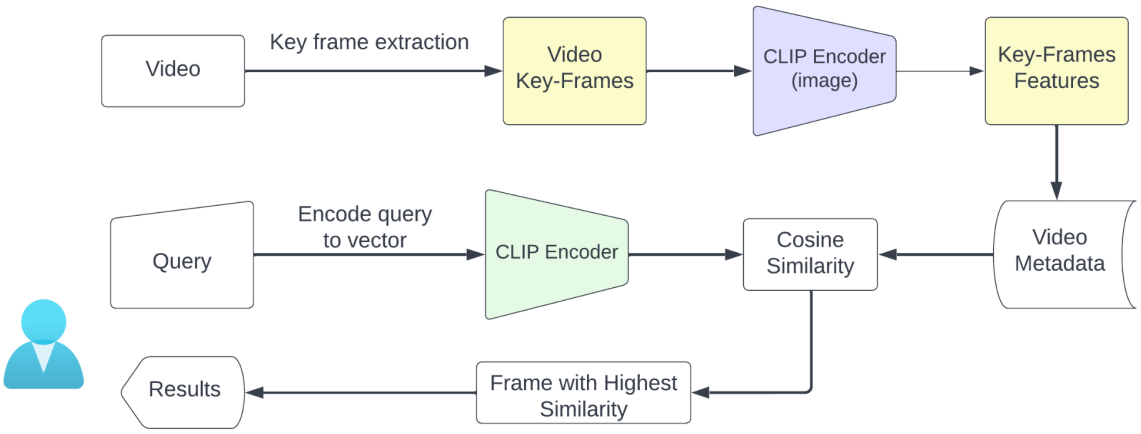

Figure 1. Overview of the Method

Method

The overview of the proposed method is shown in Figure 2. Following are the steps of the proposed method:-

[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International Conference on Machine Learning. PMLR, 2021.

- Keyframes (frames that define the starting and ending points of a smooth transition) are extracted from the video using the FFmpeg library.
- The keyframes images are encoded into a 512-dimensional vector using CLIP image encoder.
- These vectors are image features which will be stored in the storage as metadata.
- The user gives his query in text or image.
- If the query is text, CLIP text encoder is used to get the feature vector. If the query if an image, CLIP image encoder is used.
- Cosine Similarity between the query feature and the features stored as metadata is calculated and the frame(timestamp) with highest similarity is returned.

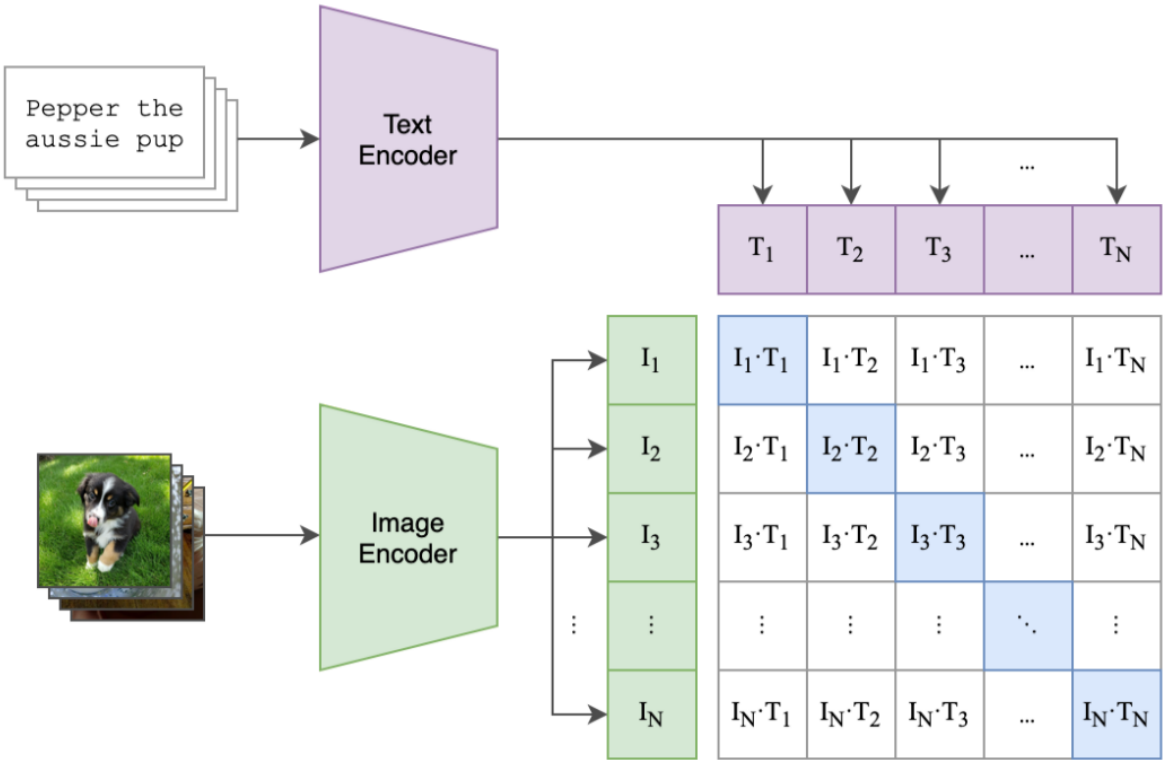

Figure 2: CLIP (Contrastive Language–Image Pre-training)

Conclusion

In this project, we have developed a system to retrieve videos using the queue. To get the semantic relationship between the query and the image features, we have used CLIP. The metadata information of the video contains the most informative parts of the video, which is later utilized for the query. The metadata information helped to get faster results for the query.