Using Vision-Language Model for Precise Frame Identification in YouTube Videos

NG Kin Pak
1155143402@link.cuhk.edu.hk

LI Yinxi
1155160255@link.cuhk.edu.hk

February 24, 2024

1 Overview

In this project, we aim to explore the potential of Vision-Language Models, such as the Contrastive Language-Image Pre-training (CLIP) model. The objective of this project is to develop a video searching system capable of identifying specific frames or moments within YouTube videos that are relevant to a given textual query to offer efficient navigation and interaction between users and video content.

2 Background

Efficiently locating and pinpointing specific moments within a sea of online video content or stock footage is one of the significant challenges in today's digital age. Traditional solutions rely on human identification of video content or various video analysis tools attempt to categorize and summarize video content based on visual and audio features. Nevertheless, both approaches, which predefine the summarized description of the content, possess constraints. The use of Vision-Language Models, such as CLIP, could improving the searchability and accessibility of video content, making it easier for users to find exactly what they're looking for without relying solely on metadata.

3 Methodology

The project will focus on the following key areas:

1. Integration of Vision-Language Model (VLM): Leveraging the pre-trained model such as CLIP, the system will be designed to process text queries and retrieve the most relevant video frames. The integration will focus on ensuring that the model's inference capabilities are finely tuned for accuracy in matching frames with text descriptions.

2. System Development: Building the core system architecture that allows users to input text queries and receive corresponding video frame outputs. This will involve developing a user interface (UI) that is intuitive and responsive, as well as backend services that handle query processing, model interaction, and video handling efficiently.

3. Performance Optimization: Implementing techniques to accelerate the response time of the system, such as optimizing the model inference pipeline, employing efficient indexing and retrieval mechanisms for video content, and utilizing hardware acceleration where possible. The goal is to minimize latency to provide real-time or near-real-time user experience.

4. Usability Enhancements: Ensuring that the system is not only fast but also easy to use. This involves designing the UI to guide users effectively through the search process and present results in a way that makes it simple to navigate to the desired video frame.

4 Technical Challenges

The primary technical challenge in this project lies in enhancing the user experience such as speed and reliability of the Vision-Language Model (VLM) beyond the existing benchmarks set by the SOTA models on this specific project. While some SOTA models such as CLIP already demonstrates robust performance, our goal is to push the boundaries further by exploring advanced optimization strategies. To this end, we plan to investigate various machine learning system enhancements such as:

- Intelligent Sampling: Developing algorithms for selective sampling of video frames that reduce the computational load without compromising the accuracy of the model.
- Repetition Reduction: Implementing mechanisms to recognize and avoid processing of repetitive content within videos, thereby increasing efficiency.
- State-of-the-Art (SOTA) Techniques: Testing and integrating the latest advancements in machine learning and computer vision to refine the model's performance, including real-time learning and adaptation to new video content types.

We anticipate that these enhancements will require a deep dive into the trade-offs between computational efficiency and model accuracy. We plan to design the machine learning system to balance this trade-off to achieve an optimal user experience.

5 Technical Impact

The successful implementation of this project is expected to have a significant impact on the field of video content management and retrieval. By showcasing the effectiveness of VLMs in real-world applications, we aim to:

- Illustrate the Versatility of VLMs: Demonstrate the wide array of practical applications for VLMs, encouraging developers to adopt and adapt these models for innovative uses.
- Propel Software Development: Provide a blueprint for integrating VLMs into software solutions, influencing future tools and platforms to leverage the full potential of vision and language understanding.
- Stimulate Research and Development: Inspire further research into optimization techniques for VLMs, potentially leading to breakthroughs that could benefit various domains such as online education, digital libraries, and content creation.

Through this project, we aim to not only advance the state of technology in video frame retrieval but also to provide a valuable reference for future research and commercial ventures in the domain of artificial intelligence and multimedia content processing.

6 Social Impact

By improving the searchability of video content, this technology can impact how information is consumed and shared with societies. There are several possible changes.

- Increased efficiency and accuracy in accessing video-based information.
- Enhanced work productivity in research or jobs related to video.
- Enhanced work productivity in research or jobs related to video.
- Crime investigation