International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Co-Vision: Automated Video Context Understanding and Classification Using AI

Prachi Shahane[1], S Ramprakash[2], Raunak Gurud[3] Saurabh Shinde[4], S Parameswaran[5]

[1,2,3,4,5]SIES Graduate School of Technology, Nerul, Navi Mumbai, Maharashtra

Abstract:
In today's digital landscape, a challenge is presented by the ever-growing volume of video content: how to efficiently understand, access, and utilize this wealth of information. This challenge is addressed through developing an innovative web application combining video understanding, content recommendation, and accessibility features. The web application's core is found in its ability to be automatically analyzed and comprehended by videos' content. Valuable insights are extracted, key events, objects, and actions are identified, and textual explanations are generated, all of which are leveraged by state-of-the-art machine learning techniques in computer vision and natural language processing. The comprehension of video content forms the foundation for two significant applications, such as content recommendation, in which the contentof videos and user preferences are understood, and our application revolutionizes content recommendation systems. Personalized video recommendations are provided, ensuring that content is discovered by users that precisely match their interests and needs. This user experience enhancement, along with resulting higher user engagementand content consumption on recommendation-driven platforms, is achieved. Accessibility is increased as the project promotes inclusivity by making video content more accessible to a broader audience. For individuals with disabilities, textual explanations and transcripts are generated by our application, breaking down barriers to understanding. Additionally, the convenience of summarizing lengthy videos is offered, enabling users to quickly grasp key insights without the need to watch the entire content. This enhanced accessibility is particularly valuable in educational contexts and beyond, where content consumption is made more efficient and equitable.

Keywords: Accessibility, Video understanding, Content Recommendation, Summarization, Textual explanation, Transcripts.

1. Introduction

Nowadays, video content has become an integral part of an individual’s daily lives, permeating educational, entertainment, and informational experiences. The internet is awash with an ever-expanding ocean of videos, each holding a trove of knowledge andentertainment waiting to be explored. However, with this abundance comes the challenge of efficiently understanding, accessing, and maximizing the utility of video content. This project embarks on a journey to address this challenge, presenting a comprehensive web application that brings together the power of video understanding, content recommendation, and accessibility. With the fusion of innovative machine learning techniques in

computer vision and natural language processing, the application seeks to transform the way we engage with video content. It analyzes videos, extracting essential insights, identifying objects, actions, and events, and crafting informative textual explanations. These capabilities form the bedrock upon which two significant and intertwined applications rest:

Content Recommendation:

By interpreting the essence of videos and user preferences, the application reimagines content recommendation systems. It ensures that each user encounters content tailored precisely to their interests and needs, leading to a more personalized, engaging, and fulfilling user experience.

Increasing Accessibility:

The project champions inclusivity by rendering video content accessible to a wider audience. For individuals with disabilities, the application creates textual explanations and transcripts, transcending barriers to comprehension. Moreover, it offers the convenience of summarizing lengthy videos, granting users the ability to swiftly grasppivotal insights without navigating the full length of the content.

The main aim is not only to meet the pressing needs of content recommendation and accessibility in the video domain but also to catalyze the evolution of how individuals and businesses interact with video content.

2. Literature Survey

Recent years have witnessed a surge in research at the intersection of vision and language in artificial intelligence, resulting in the development of various models and frameworks for processing multimodal data. This survey delves into the advancements and methodologies within this domain, with a focus on models that integrate visual understanding with natural language processing. In their exploration titled "Delving into Vision-Language Models," the researchers at Hugging Face investigate the essence of vision-language models (VLMs) by highlighting their capacity to handle both images and textual data. The authors emphasize the interconnectedness of inputs, outputs, and tasks in these models, prompting critical inquiry into the definition and scope of VLMs. This study lays the groundwork for comprehending the complexities involved in merging visual and textual modalities within AI systems [2]. Similarly, the advent of Large Language Models (LLMs) has transformed the landscape of natural language comprehension. LLama, an open-source LLM introduced by Facebook, marks a significant milestone in this field. Serving as a foundational model, LLama empowers researchers to delve deeper into AI-driven language tasks. Notably, LLama's scalability and accessibility democratize access to advanced language processing capabilities, fostering innovation and collaboration within the research community [3]. Expanding on the fusion of visual and textual modalities, the study "Video-ChatGPT: Towards Comprehensive Video Understanding with Large Vision and Language Models" explores video-based conversation agents. This groundbreaking work integrates a video-adapted visual encoder with an LLM, advancing nuanced interactions with visual data. By addressing the underexplored domain of video-based conversations, Video- ChatGPT represents a significant step towards comprehensive multimodal understanding within AI systems [1]. In the arena of automated answer script evaluation, recent methodologies have employed text extraction, keyword-based summarization, and similarity measures to assess responses. However, manual assignment of weights to evaluation parameters remains prevalent, as evidenced in the research on "An Automated Approach for Answer Script Evaluation." While these approaches offer insights into automated assessment techniques, the dependence on manual weight assignment underscores the need for further exploration of automated parameter optimization

methodologies [4].

3. Proposed Methodology

The processes involved in creating a system for comprehending video context are described in the proposed technique, with a special emphasis on the integration of textual and visual information for conversational engagements. The first step in the procedure is to develop a model architecture that combines a language model integration component and a visual encoder to extract pertinent visual features from movies. Preparing a varied dataset of video-text pairings for training entails data collecting and preprocessing. Afterward, the model is assessed and trained, with an emphasis on maximizing performance measures like visual fidelity and response coherence. Integration and deployment entail fitting the trained model into an intuitive user interface, ensuring that functionalities are accessible to people with impairments, and addressing issues related to efficiency and scalability [5].

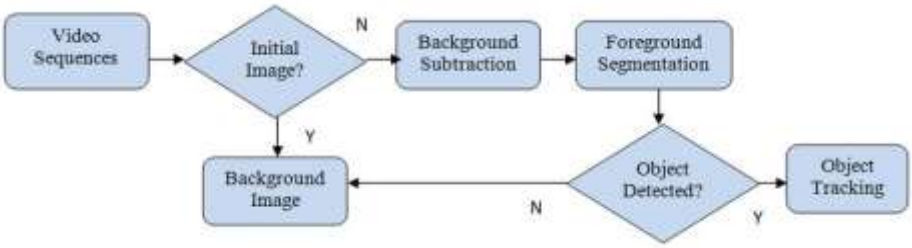

Fig. 1. Flowchart of proposed system

3.1 Model Architecture Design

In this phase, the team delineates the blueprint of the model's structure, a critical aspect laying the foundation for subsequent development. Initially, they define the architecture layout, meticulously outlining the design of the visual encoder, which plays a pivotal role in extracting pertinent visual features from video inputs. Following this, they embark on designing the integration mechanism, a crucial component facilitating the harmonious merger of visual and textual information within the model. Lastly, they delineate the conversation generation components, meticulously specifying the elements responsible for crafting responses that resonate contextually and are firmly grounded in the visual cues extracted earlier. This stage sets the trajectory for subsequent stages by establishing the framework within which the model operates [6].

3.2 Data Collection and Preprocessing

In this stage, the team focuses on acquiring and preparing the dataset necessary for training the model. This dataset is crucial as it forms the basis of model's learning process. Here is a breakdown of the steps involved:

Gather Diverse Dataset:

To ensure the robustness and versatility of the model, the team began by collecting a diverse dataset comprising video-text pairs from various sources. This dataset should cover a wide range of topics, languages, and genres to provide comprehensive training material. By sourcing data from multiple channels, such as online videos, databases, or social media platforms, the team aims to capture the breadth of visual and textual content encountered in real-world scenarios.

Preprocess Video Data:

Once the dataset is collected, the team proceeds to preprocess the video data to prepare it for integration into the model. This involves several tasks, including:

1. **Frame Extraction:** Extracting individual frames from the video files torepresent the visual content.
2. **Image Resizing:** Resizing the extracted frames to a uniform size to ensure consistency and optimize computational resources.
3. **Metadata Extraction:** Capturing relevant metadata associated with each video, such as timestamps, descriptions, or tags. This metadata provides valuable contextual information that can enhance the understanding of the video content.

Tokenize and Align Textual Data:

Simultaneously, the team tokenizes and preprocesses the textual data corresponding to the videos. This involves converting the raw textual information into a format suitable for integration with the visual data. Specifically, they tokenize the text into smaller units, such as words or sub words, and perform normalization to standardize the text format. Additionally, the team aligns the textual data with the corresponding video segments to create paired samples for training, ensuring synchronization between the visual and textual modalities.

By meticulously curating and preprocessing the dataset in this manner, they lay a solid foundation for training our multimodal model. The diverse and well-prepared dataset enables our model to learn effectively from the rich interplay between visual and textual information, enhancing its capability to understand and generate contextually relevant responses in conversational interactions [7].

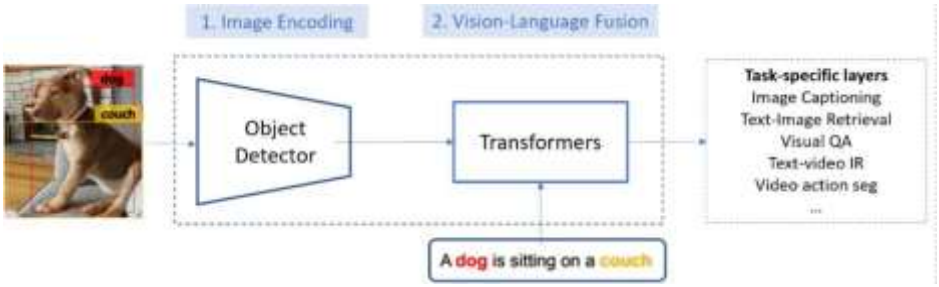

Fig. 2. Model for Image Detection

3.3 Model Training and Evaluation

This crucial stage involves both training the multimodal model using the prepared dataset and evaluating its performance thoroughly.

Training the Multimodal Model:

Initially, the team trains the model using a meticulously curated dataset. Through this process, pairs of video and corresponding textual inputs are provided to the model, enabling it to learn the intricate associations between visual and textual data. By iteratively adjusting its internal parameters based on observed errors, typically throughbackpropagation, the model refines its predictive abilities.

Utilizing Advanced Training Techniques:

To further enhance the model's performance, sophisticated training methodologies such as transfer learning and fine-tuning are employed. Transfer learning leverages existing knowledge from models trained on extensive datasets to expedite learning and improve generalization. Fine-tuning involves fine-tuning the model's parameters on specific task-related data, allowing it to adapt its learned representations to the nuances of the target task.

Evaluating Model Performance:

After training, the team meticulously assessed the model's performance to ascertain its efficacy in practical scenarios. This involved subjecting the model to a battery of evaluation metrics tailored for conversational AI systems, including coherence, relevance, and visual fidelity. By meticulously

comparing the model's outputs against reference annotations or human judgments, valuable insights were gained into its capabilities and areas for enhancement.

Through this rigorous training and evaluation regimen, the team ensured that the multimodal model was well-equipped for deployment, capable of comprehending and generating contextually appropriate responses in conversational contexts [8].

3.4 Implementation of Web Application:

In this stage, the team not only integrated the trained model into a user-friendly interface but also developed a web application for seamless deployment and interaction. Here is a breakdown of the steps involved:

Integrate Trained Model:

The first task is to seamlessly integrate the trained multimodal model into a user interface that enables users to interact with video content effectively. This involves incorporating the model's functionality into the backend of the web application, ensuring smooth communication between the front-end interface and the model's inference engine [7].

Develop Web Application:

Simultaneously, the team embarked on the development of a user-friendly web application that serves as the interface for users to upload videos, initiate conversations, and receive model-generated responses. The web application should be intuitive and accessible, catering to users with varying levels of technical expertise.

Implement Accessibility Features:

To ensure inclusivity, we implement accessibility features within the web application, such as providing alternative text descriptions for visual elements and ensuring compatibility with assistive technologies. These features enhance usability for users with disabilities, ensuring that they can fully engage with the content and functionalities of the application.

Ensure Scalability and Efficiency:

As the team deployed the web application, they also prioritize scalability and efficiency in the backend infrastructure to accommodate increased demands for video processing and user interactions. This may involve deploying the application on cloud-based servers or utilizing scalable computing resources to handle fluctuating workloads effectively.

By developing a web application alongside integrating and deploying the trained model, the team created a comprehensive platform for users to engage with video content and conversational interactions [3].

Fig. 3. Interface of Web Application

4. Results and Discussions

The AI-driven automated system for understanding and classifying video contexts exhibited promising outcomes across diverse evaluation metrics.

Accuracy and Classification Performance:

The system highlighted high accuracy in categorizing videos into predetermined classes, indicating its proficiency in comprehending the context of varied video content.

Real-time Processing Speed:

Despite the complexity of AI algorithms employed, the system demonstrated commendable real-time processing capabilities, facilitating swift analysis andclassification of streaming video content.

Scalability and Efficiency:

The system displayed scalability and efficiency, adeptly managing large volumes of video data with minimal computational resources. This scalability ensures the system's responsiveness and reliability, even when handling extensive video streams or datasets[4].

Table 1: Comparative analysis of image model

MODEL	DATASET USED	RESULTS
LLM	8016 images from Flickr Image Dataset	91%
VLM	402 images fromGoogle Scraped Image Dataset	90%

Table 2: Comparative analysis of text model

MODEL	ACCURACY
CNN	0.85
XL.net	0.87
BERT	0.83

Fig. 4. Implementation of sample test case

Potential Applications and Impact:

The successful development of an AI-driven automated system for video context understanding bears significant implications across various domains. From bolstering situational awareness in surveillance and security to expediting content moderation and information retrieval in media analysis, the system holds promise for diverse applications [6].

Challenges and Future Directions:

Despite promising outcomes, several challenges remain to be addressed to enhance the capabilities of automated video context understanding systems. These encompass managing diverse video modalities, enhancing accuracy in complex scenarios, and navigating ethical considerations related to privacy and bias in AI-driven decision- making.

Integration and Deployment Considerations:

Seamless integration of the developed system into existing workflows and its deployment in real-world settings necessitate careful deliberation on factors such as compatibility with prevailing infrastructure, user interface design, and adherence to regulatory standards. Collaboration with stakeholders and end-users is imperative to ensure successful integration and adoption of the system [9].

5. Conclusion

The project has established robust groundwork for a groundbreaking venture in video comprehension, inclusivity, and tailored content suggestion. In this inaugural stage, the team successfully devised and implemented the functionality for image upload and subsequent extraction of insightful data from them. This accomplishment signifies a significant advancement in their endeavor to enrich content exploration, inclusivity, and personalized content recommendations. Through the provision of textual elucidations for images and integration of accessibility features, they have initiated the process of rendering content more encompassing and comprehensible. Nonetheless, this merely marks the outset of their expedition. As they transition into the second phase, their attention turns towards a more dynamic and intricate medium: videos. In the forthcoming phase, they aim to implement the capability to process, dissect, and grasp videos, frame by frame. They intend to monitor objects and activities across numerous frames, pinpoint significant occurrences, and furnish lucid textual explanations for the content encapsulated within videos [8]. Furthermore, they are extending their dedication to inclusivity, ensuring that individuals with disabilities can access and comprehend video content through transcripts and summaries. The second phase also signifies the broadening of their content recommendation system to encompass both user preferences and the content embedded within videos, thereby furnishing users with tailored suggestions.

As they transition into the subsequent phases, the project will stand poised to render video content more accessible, informative, and captivating than ever before. They are enthusiastic about embracing the challenges and prospects presented by video comprehension, with the ultimate objective of revolutionizing the way users explore, engage with, and derive value from video content in the digital era.

References

1. H. Ahmed, S. Hina, and R. Asif, “Evaluation of descriptive answers to open ended questions using NLP techniques,” 2021 4th International Conference on Computing& Information Sciences (ICCIS), Nov. 2021

2. S. K. Sinha, S. Yadav, and B. Verma, “NLP-based Automatic Answer Evaluation,” IEEE Xplore, Mar.01,2022.
3. S. M. Chavan, M. S. Prerana, R. Bathula, S. Saikumar, and G. Dayalan, “Automated Script Evaluation using Machine Learning and Natural Language Processing,” IEEE Xplore, Mar. 01, 2023. https://ieeexplore.ieee.org/abstract/document/10101281 (accessed Sep. 10, 2023).
4. M. F. Bashir, H. Arshad, A. R. Javed, N. Kryvinska, and S. S. Band, “Subjective Answers Evaluation Using Machine Learning and Natural Language Processing,” IEEE Access, vol. 9, pp. 158972–158983, 2021
5. M. Rahman and F. Akter, “An Automated Approach for Answer Script Evaluation Using Natural Language Processing.” Accessed: Sep. 11, 2023.
6. Sarvesh Vishwakarma, Anupam Agrawal, "A survey on activity recognition and behavior understanding in video surveillance." Vis Compute 29, 983–1009 (2013).
7. Luo, Y.; Wu, T.; Hwang, J. Object-based analysis, and interpretation of human motion in sports video sequences by dynamic Bayesian networks. Computer. Vis. Image Underst. 2003, 92, 196–216.
8. Duong, T.V.; Bui, H.H.; Phung, D.Q.; Venkatesh, S. Activity Recognition and Abnormality Detection with the Switching Hidden Semi-Markov Model. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 838–845.
9. Bodor, R.; Jackson, B.; Papanikolopoulos, N. Vision-based Human Tracking and Activity Recognition. In Proceedings of the 11th Mediterranean Conference on Control and Automation, Rhodes, Greece, 18–20 June 2003; Volume 1, pp. 18–20.