

MỘT PHƯƠNG PHÁP TÍCH HỢP CHO BÀI TOÁN HỎI ĐÁP TRÊN VIDEO CÓ ĐỘ DÀI LỚN VỚI HIỂU BIẾT VỀ HÌNH ẢNH VÀ ÂM THANH

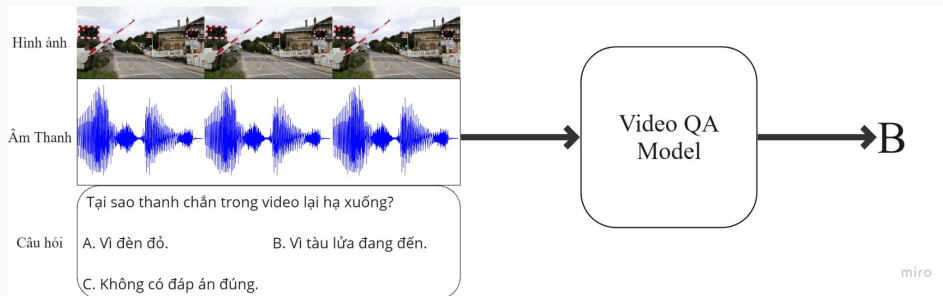
Võ Anh Hào - 21520832

Tóm tắt



- Lớp: CS519.011
- Link Github của nhóm:
- Link YouTube video:
- Họ và Tên: Võ Anh Hào

Giới thiệu



Input và Output của bài toán

Examples of Video QA tasks and their corresponding outputs:

Example 1:

User: Summarize the plot of this movie.

Assistant: Avatar follows Jake Sully, a disabled former Marine who is recruited by the RDA to work on Pandora, a moon rich in a valuable mineral called unobtainium. Jake's twin brother, Tommy, died in a helmet-cam accident, and Jake is offered his place in the Avatar Program, which allows humans to link with the native Na'vi, a humanoid species. Jake is assigned to the avatar of his deceased brother....

Example 2:

User: How human work on Pandora?

Assistant: Humans work on Pandora by using avatars, which are grown from human DNA mixed with the DNA of the native Na'vi, to interact with the environment and the Na'vi.

Example 3:

User: Why Jake Sully decide to fight against humanity with the Na'vi?

Assistant: Jake Sully decides to fight against humanity with the Na'vi after he falls in love with the Na'vi and their way of life, and after he realizes that the human military is destroying their home and culture.

Example 4:

User: Describe what you hear?

Assistant: I hear sounds of footsteps on the floor in the background, and there is also a dog barking in the house.

User: Does the man wear glasses?

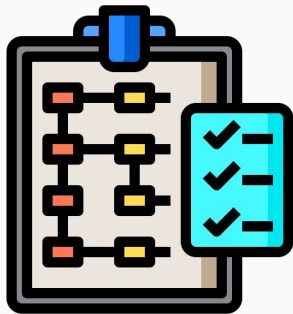
Assistant: Yes, he wears a pair of glasses in the video.

Một vài ví dụ của các nghiên cứu hiện nay

Mục tiêu



Xây dựng mô hình hỏi đáp trên video có độ dài lớn, đồng thời có khả năng tích hợp thông tin âm thanh và hình ảnh



Thực hiện đánh giá và so sánh với các phương pháp hiện tại trên các bộ dữ liệu thích hợp

Nội dung và Phương pháp

Nội dung:

- Nghiên cứu các phương pháp để tích hợp thông tin âm thanh vào mô hình từ các module tiền huấn luyện.
- Tìm hiểu các kỹ thuật huấn luyện khác nhau cho bài toán đặc biệt là với việc hỏi đáp trên video có độ dài lớn.
- Tiến hành xây dựng mô hình và thực hiện các thực nghiệm, đánh giá và so sánh giữa các phương pháp trước.

Nội dung và Phương pháp

Phương pháp:

- Nghiên cứu các kiến trúc và phương pháp như Video-LLaMA [2], ImageBind [3], AudioGPT [4] từ đó ứng dụng để tích hợp thông tin âm thanh vào mô hình.
- Tìm hiểu kiến trúc và phương pháp huấn luyện của LLaMA-VID [1] để áp dụng cho quá trình huấn luyện mô hình: điều chỉnh tương thích (Modality Alignment), điều chỉnh hướng dẫn (Instruction Tuning) và đặc biệt là điều chỉnh tương thích với video có độ dài lớn(Long Video Tuning).
- Thực nghiệm tích hợp các mô hình trích xuất thông tin âm thanh vào mô hình LLaMA-VID [1], huấn luyện mô hình với các kiến trúc và kĩ thuật khác nhau sau đó so sánh và đánh giá trên các tập dữ liệu trả lời câu hỏi nghe nhìn (Audio-Visual Question Answering) trên video có độ dài khác nhau như: AVQA[5], MovieQA[6].

Kết quả dự kiến

- Tài liệu báo cáo các phương pháp, cách tiếp cận liên quan và các kỹ thuật huấn luyện.
- Tài liệu báo cáo về mô hình đề xuất và kết quả thực nghiệm đánh giá mô hình đề xuất với các mô hình khác trên các bộ dữ liệu trả lời câu hỏi nghe nhìn (Audio-Visual Question Answering).

Tài liệu tham khảo

- [1]. Yanwei Li, Chengyao Wang, Jiaya Jia: LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models. CoRR abs/2311.17043 (2023)
- [2]. Hang Zhang, Xin Li, Lidong Bing: Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. EMNLP (Demos) 2023: 543-553
- [3]. Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, Ishan Misra: ImageBind One Embedding Space to Bind Them All. CVPR 2023: 15180-15190
- [4]. Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Zhou Zhao, Shinji Watanabe: AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head. CoRR abs/2304.12995 (2023)
- [5]. Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, Wenwu Zhu: AVQA: A Dataset for Audio-Visual Question Answering on Videos. ACM Multimedia 2022: 3480-3491
- [6]. Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, Sanja Fidler: MovieQA: Understanding Stories in Movies through Question-Answering. CVPR 2016: 4631-4640