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

Trường Đại học Công nghệ Thông tin, ĐHQG-HCM

Giới Thiệu

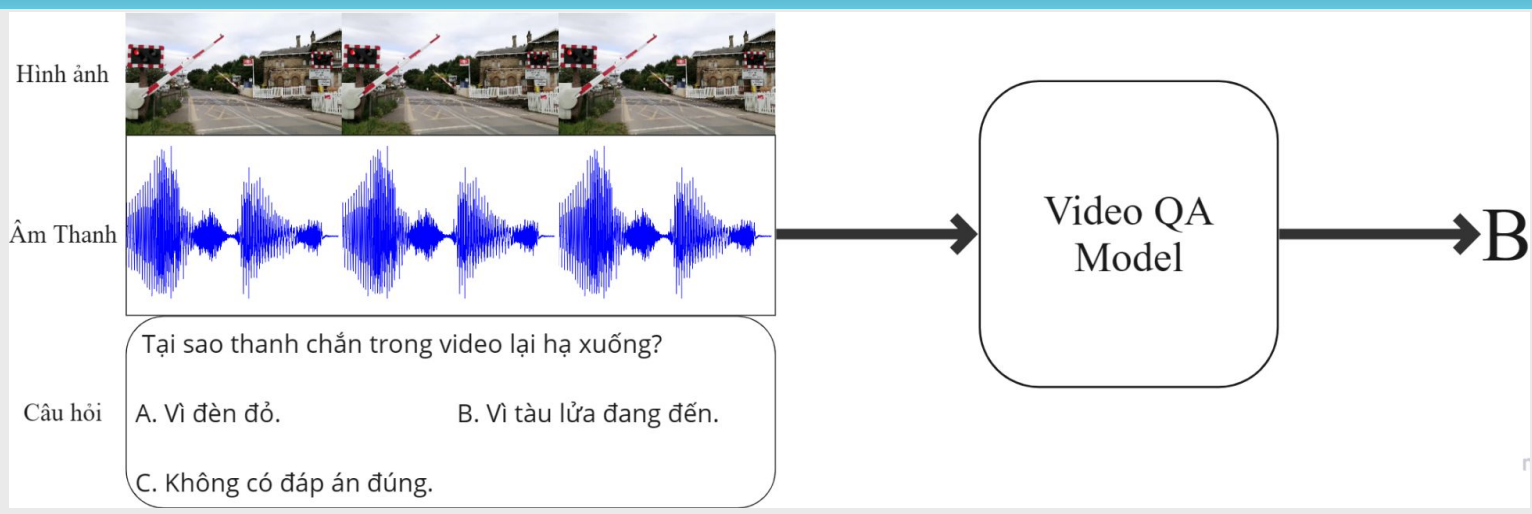
Chúng tôi mong muốn phát triển một phương pháp mới cho bài toán hỏi đáp dựa trên nội dung video, cụ thể:

- Dựa trên hai nghiên cứu LLaMA-VID và Video-LLaMA để làm tiền đề xây dựng một mô hình có khả năng hỏi đáp trên video có độ dài hàng giờ đồng thời tích hợp thông tin âm thanh và hình ảnh.
- Sau đó, đánh giá trên các tập dữ liệu khác nhau như: AVQA, MovieQA.

Động Lực

- **Khả năng ứng dụng:** Áp dụng trong nhiều lĩnh vực khác nhau, từ giải trí đến giáo dục và nghiên cứu.
- **Bối cảnh:** Các mô hình hiện tại chưa hiệu quả trong việc tích hợp thông tin âm thanh trong video dài.
- **Tầm Quan Trọng của Âm Thanh:** Âm thanh cung cấp thông tin bổ sung quan trọng cho nội dung, cảm xúc và bối cảnh.

Tổng Quan



Hình 1. Đầu vào và đầu ra của bài toán

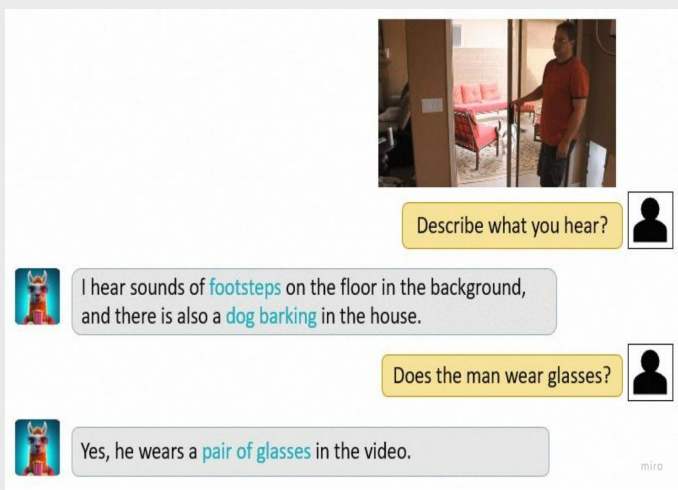
Mô Tả

1. Nội dung

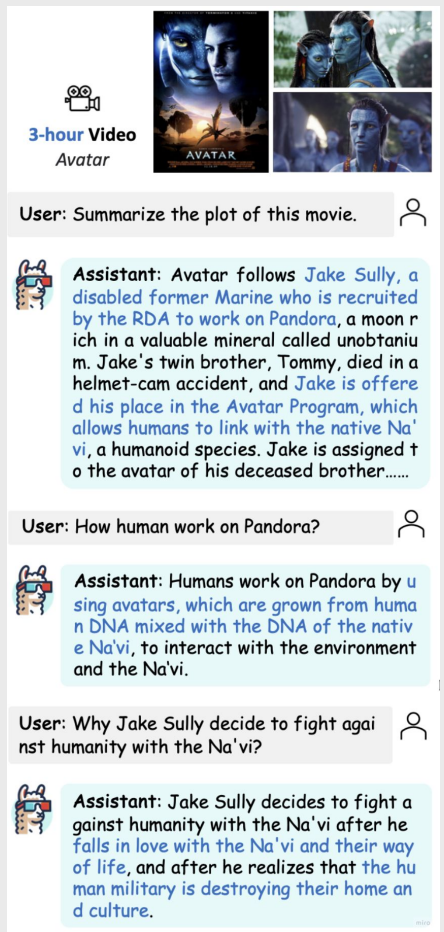
- Nghiên cứu các phương pháp để tích hợp thông tin âm thanh vào mô hình từ các module tiền huấn luyện.
- Tìm hiểu các kỹ thuật huấn luyện khác nhau cho bài toán đặc biệt là với việc hỏi đáp trên video có độ dài lớn.
- Tiến hành xây dựng mô hình và thực hiện các thực nghiệm, đánh giá và so sánh giữa các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

2. Phương pháp

- Nghiên cứu các kiến trúc và phương pháp như Video-LLaMA (Hình 2), ImageBind, AudioGPT 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 (Hình 3) 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 dài (Long Video Tuning).



Hình 2. Ví dụ về khả năng tích hợp thông tin âm thanh của Video-LLaMA



Hình 3. Ví dụ về khả năng trả lời câu hỏi trên video có độ dài hàng giờ của LLaMA-VID

NII

Tên tác giả chính– Trường Đại học

TEL : ******* Email : *********