


THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
https://youtu.be/Z_rz3t_-m9w
- Link slides (dạng .pdf đặt trên Github của nhóm):
<https://github.com/hao3830/CS519.O11/blob/master/document.pdf>
- Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới
- Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in

<ul style="list-style-type: none">● Họ và Tên: Võ Anh Hào● MSSV: 21520832 	<ul style="list-style-type: none">● Lớp: CS519.O11● Tự đánh giá (điểm tổng kết môn): 9/10● Số buổi vắng: 1● Số câu hỏi QT cá nhân: 11● Số câu hỏi QT của cả nhóm: 11● Link Github: https://github.com/mynameuit/CS519.O11/
---	---

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

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

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

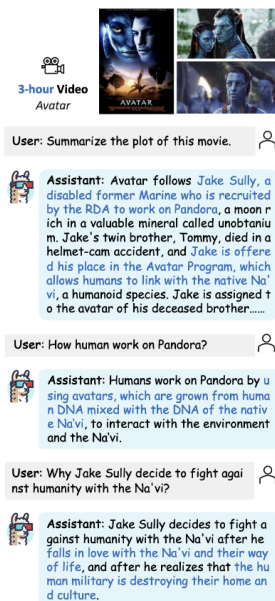
AN INTEGRATED APPROACH FOR VIDEO QUESTION ANSWERING ON LONG-DURATION VIDEOS WITH VISUAL AND AUDIO CONTENT

TÓM TẮT (Tối đa 400 từ)

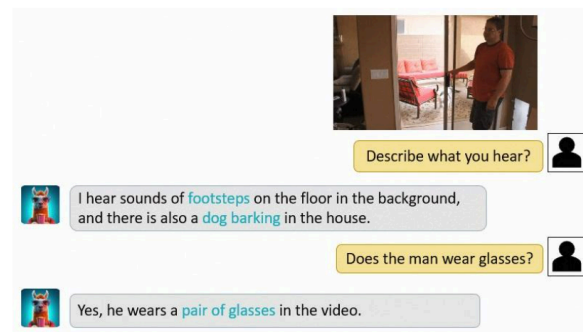
Trong những năm gần đây, việc tiêu thụ sản phẩm video đã trải qua sự bùng nổ chưa từng thấy trên các nền tảng khác nhau như Youtube, Netflix và các phương tiện truyền thông xã hội, tạo ra nhu cầu ngày càng tăng về các kỹ thuật hiểu và truy xuất video hiệu quả đặc biệt là với cái video có độ dài lớn như phim ảnh. Để giải quyết nhu cầu trên, trả lời câu hỏi dựa trên video (Video Question Answering) đã nổi lên như một bài toán quan trọng trong lĩnh vực thị giác máy tính và xử lý ngôn ngữ tự nhiên. Hai phương pháp nghiên cứu gần đây, LLaMA-VID [1] và Video-LLaMA [2], đã cung cấp những nền tảng quan trọng cho việc giải quyết bài toán này. Tuy nhiên, cả hai phương pháp này vẫn còn thiếu sự tích hợp hoàn hảo của thông tin âm thanh trong quá trình xử lý video có độ dài lớn. Vì vậy, trong nghiên cứu này chúng tôi mong muốn kết hợp các ý tưởng từ hai phương pháp trên nhằm phát triển một mô hình có khả năng trả lời câu hỏi dựa trên nội dung của video có độ dài lớn, đồng thời tận dụng hiệu quả các thông tin hình ảnh và âm thanh. Dự án nghiên cứu sẽ bao gồm việc nghiên cứu các kỹ thuật để tích hợp các mô hình rút trích thông tin âm thanh tiền huấn luyện vào mô hình LLaMA-VID [1] - một mô hình có khả năng hỏi đáp dựa trên hình ảnh của video có độ dài hàng giờ đồng hồ, khám phá các chiến lược đào tạo khác nhau cho bài toán và đánh giá mô hình trên các bộ dữ liệu trả lời câu hỏi nghe nhìn với độ dài video đa dạng, chẳng hạn như AVQA [5] và MovieQA [6].

GIỚI THIỆU (Tối đa 1 trang A4)

Trong những năm gần đây, việc tiêu thụ sản phẩm video đã trải qua sự bùng nổ chưa từng thấy trên các nền tảng khác nhau như Youtube, Netflix và các phương tiện truyền thông xã hội. Do đó, nhu cầu về các kỹ thuật hiểu và truy xuất video hiệu quả ngày càng trở nên cấp thiết. Trả lời câu hỏi dựa trên video (Video Question Answering) đã phát triển như một bài toán tối quan trọng trong lĩnh vực thị giác máy tính (Computer Vision) và xử lý ngôn ngữ tự nhiên (NLP).



(a) Ví dụ về khả năng của LLaMA-VID khi thực hiện hỏi đáp với video có độ dài lớn



(b) Ví dụ về khả năng của Video-LLaMA khi thực hiện hỏi đáp về thông tin âm thanh và thông tin hình ảnh của video

Hình 1. Một vài ví dụ về khả năng hỏi đáp trên video của các nghiên cứu hiện nay

Video Question Answering (Video QA) là một bài toán đa nhiệm phức tạp, yêu cầu hệ thống phải có khả năng hiểu và trả lời câu hỏi dựa trên nội dung của video. Điều này không chỉ liên quan đến việc phân tích hình ảnh và văn bản xuất hiện trong video mà còn bao gồm cả thông tin âm thanh. Các nghiên cứu gần đây [1][2][4] đã cải thiện đáng kể hiệu năng của các mô hình trên bài toán Video QA, nhờ vào việc kết hợp với mô hình ngôn ngữ lớn (LLMs) từ đó tạo ra khả năng diễn giải và suy luận tốt hơn cho hệ thống. Ví dụ, mô hình LLaMA-VID [1] (Hình 1a) cho phép thực hiện hỏi đáp trên video với độ dài hàng giờ đồng hồ nhờ vào việc tối ưu chi phí tính toán bằng cách chỉ sử dụng hai token bao gồm: context token và content token để biểu diễn thông tin cho

mỗi khung hình trước khi đưa vào mô hình LLMs để xử lý. Tuy nhiên, cách tiếp cận này lại thiếu đi việc khai thác thông tin âm thanh của video, một trong những thành phần mang nhiều thông tin quan trọng khi xử lý các video có độ dài lớn như phim ảnh. Một cách tiếp cận khác, Video-LLaMA [2] (Hình 1b) tích hợp bộ mã hóa hình ảnh và âm thanh đã được huấn luyện trước với LLMs đóng băng để đào tạo kết hợp các phương tiện (cross-modal). Tuy nhiên, phương pháp này lại gặp hạn chế với số lượng khung hình xử lý mỗi lần hạn chế, ảnh hưởng đến khả năng xử lý video dài.

Vì vậy, trong đề tài này, tôi sẽ nghiên cứu kết hợp hai cách tiếp cận trên [1][2] cũng như tham khảo các phương pháp khác cho bài toán video QA [4], từ đó xây dựng một mô hình có khả năng thực hiện hỏi đáp dựa trên video có độ dài lớn cũng như có khả năng khai thác cả thông tin âm thanh và hình ảnh của video. Như vậy, đầu vào và đầu ra của bài toán như sau:

- Đầu vào: Câu hỏi của người dùng và đoạn video bao gồm cả hình ảnh và âm thanh.
- Đầu ra: Một câu trả lời có liên quan đến câu hỏi của người dùng dưới dạng văn bản được tạo ra dựa trên thông tin được thu thập và phân tích từ video.



Hình 2. Đầu vào và đầu ra của bài toán

MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

- Tìm hiểu các phương pháp liên quan cũng như áp dụng ý tưởng của hai phương pháp LLaMA-VID [1] và Video-LLaMA [2] để tạo ra mô hình có khả năng hỏi

đáp trên video có thời lượng hàng giờ đồng hồ bằng cách khai thác cả thông tin văn bản, hình ảnh và âm thanh.

- Thực hiện đánh giá và so sánh với các mô hình khác 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].

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

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.

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 dài (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 kỹ 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 MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

- 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 (*Định dạng DBLP*)

[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

