BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn học

CS519 - PHƯƠNG PHÁP LUẬN NGHIÊN CỨU KHOA HỌC

Lớp học

CS519.Q11

Giảng viên

PGS.TS. LÊ ĐÌNH DUY

Thời gian

09/2025 - 12/2025

----- Trang này cố tình để trống -----

THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
  (ví dụ: https://www.youtube.com/watch?v=AWq7uw-36Ng)
- Link slides (dạng .pdf đặt trên Github của nhóm):
  (ví dụ: https://github.com/mynameuit/CS519.O21.KHTN/TenDeTai.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*

• Họ và Tên: Phạm Huỳnh Long Vũ
• MSSV: 23521813
 
	• Lớp: CS519.P11
	• Tự đánh giá (điểm tổng kết môn): 9.8/10
	• Số buổi vắng: 0
	• Số câu hỏi QT cá nhân: 9
	• Số câu hỏi QT của cả nhóm: 2
	• Link Github: https://github.com/23521813/CS519.P11
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Tìm bài toán, độ đo.
	  ○ Có đóng góp ý tưởng giải pháp.
	  ○ Viết phần giới thiệu bài toán, nội dung và giải pháp.
	  ○ Trình bày video youtube.

• Họ và Tên: Nguyễn Huy Phước
• MSSV: 23521234
 
	• Lớp: CS519.P11
	• Tự đánh giá (điểm tổng kết môn): 9.8/10
	• Số buổi vắng: 0
	• Số câu hỏi QT cá nhân: 9
	• Số câu hỏi QT của cả nhóm: 2
	• Link Github: https://github.com/23521813/CS519.P11
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Lên ý tưởng thuật toán, thiết kế độ do.
	  ○ Viết phần nội dung phương pháp, kết quả mong đợi.
	  ○ Trình bày video YouTube.

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

TRUY HỒI KHOẢNH KHẮC TRONG VIDEO

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

VIDEO CORPUS MOMENT RETRIEVAL

TÓM TẮT *(Tối đa 400 từ)*

Đề tài tập trung nghiên cứu bài toán *Video Corpus Moment Retrieval*, nhằm xây dựng một hệ thống nhận vào một mô tả bằng ngôn ngữ tự nhiên và một tập video lớn, sau đó vừa lựa chọn được video liên quan, vừa xác định khoảng thời gian trong video phù hợp với mô tả. Bài toán này đặt ra nhiều thách thức về mặt mô hình và giải pháp, đặc biệt trong bối cảnh khoảnh khắc cần tìm thường rất ngắn so với toàn bộ video và bị ẩn trong lượng dữ liệu hình ảnh khổng lồ.

Trong khuôn khổ môn Phương pháp luận nghiên cứu khoa học, đề tài hướng đến ba mục tiêu chính: (1) hệ thống hóa và phát biểu bài toán một cách rõ ràng, chỉ ra các yếu tố đầu vào, đầu ra và các tiêu chí đánh giá; (2) khảo sát các hướng tiếp cận tiêu biểu, bao gồm mô hình hai giai đoạn (truy hồi video trước, định vị khoảnh khắc sau) và mô hình end-to-end dựa trên học biểu diễn video–ngôn ngữ; (3) đề xuất một khung mô hình baseline khả thi kèm theo kế hoạch thực nghiệm.

Kết quả mong đợi là trình bày được bối cảnh bài toán, mục tiêu nghiên cứu, các thách thức chính, cùng với hướng giải pháp đề xuất và phương pháp đánh giá làm nền tảng cho những nghiên cứu tiếp theo.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Dưới sự bùng nổ của các nền tảng video (YouTube, TikTok, camera giám sát, v.v.),

lượng nội dung được tạo ra và tải lên mỗi ngày ngày càng khổng lồ. Điều này khiến việc tìm kiếm lại đúng đoạn video mà người dùng quan tâm trở nên khó khăn hơn rất nhiều. Các hệ thống tìm kiếm truyền thống thường chỉ dựa vào tiêu đề, từ khóa hoặc mô tả ngắn do người đăng tự điền, nên thường chỉ giúp tìm được "video nào đó liên quan", chứ không chỉ ra được chính xác khoảnh khắc cần xem trong một video dài. Ví dụ, với truy vấn: "người đàn ông nhảy xuống hồ rồi bơi về phía chiếc thuyền", người dùng hiện nay thường phải tự tua đi tua lại trong video để tìm đúng đoạn mình cần.

Bài toán **Video Corpus Moment Retrieval** được đặt ra để giải quyết vấn đề trên. Cụ thể, từ một mô tả bằng ngôn ngữ tự nhiên và một tập video lớn, hệ thống cần vừa xác định được video phù hợp, vừa dự đoán được khoảng thời gian trong video khớp với mô tả. Bài toán này khó ở chỗ: số lượng video trong kho có thể rất lớn; độ dài mỗi video là tùy ý, có thể lên đến hàng chục phút hoặc hàng giờ; khoảnh khắc liên quan đến truy vấn thường chỉ chiếm một phần rất nhỏ so với toàn bộ video.

Đề tài lựa chọn nghiên cứu hướng này vì mang ý nghĩa thực tiễn cao trong các ứng dụng như tìm kiếm nội dung học tập, tra cứu video giám sát, quản lý kho dữ liệu multimedia, đồng thời gắn liền với nhiều hướng nghiên cứu hiện đại về truy vấn trong video.

Cụ thể, trong đề tài này, bài toán được phát biểu như sau:

Đầu vào (Input) gồm hai thành phần chính:

1. Truy vấn (query): một câu mô tả hoặc một chuỗi sự kiện bằng ngôn ngữ tự nhiên, diễn tả một *khoảnh khắc* hoặc *chuỗi khoảnh khắc* trong video.
    - Ví dụ:
        - "Người đàn ông nhảy xuống hồ rồi bơi về phía chiếc thuyền."

- Hoặc chuỗi sự kiện: “con gà ăn thóc, sau đó con gà uống nước, rồi con gà đi ngủ”.

2. Bộ sưu tập video (video gallery): một tập gồm nhiều video có độ dài bất kỳ.

Đầu ra (Output):

1. Lựa chọn được video phù hợp nhất với truy vấn từ trong video gallery.
2. Xác định khoảng thời gian bắt đầu và kết thúc của sự kiện đó trong video tìm ra được ở bước 1.

MỤC TIÊU

Mục tiêu của đề tài là nghiên cứu một cách hệ thống bài toán Video Corpus Moment Retrieval và đề xuất một khung giải pháp cụ thể có thể triển khai và đánh giá được, với định hướng ưu tiên tiết kiệm chi phí huấn luyện và rút ngắn thời gian truy xuất kết quả so với các mô hình end-to-end phức tạp.

Tiếp theo, đề tài tiến hành khảo sát và tổng hợp các hướng tiếp cận tiêu biểu trong lĩnh vực này. Thông qua việc phân tích, đề tài hướng đến việc rút ra những ưu điểm, hạn chế.

Cuối cùng, đề tài đặt mục tiêu thiết kế và hiện thực một mô hình baseline để biểu diễn truy vấn và nội dung video, đồng thời xây dựng pipeline truy hồi và định vị nhẹ hơn về mặt huấn luyện, truy vấn nhanh.

Trước hết đề tài xác định các tiêu chí đánh giá phù hợp cho bài toán gồm độ chính xác truy hồi video và độ trùng khớp giữa khoảng thời gian dự đoán với khoảng thời gian gán nhãn.

NỘI DUNG VÀ PHƯƠNG PHÁP

Đầu tiên, đề tài tiến hành khảo sát hai phương pháp tiếp cận nổi bật trong lĩnh vực: phương pháp hai giai đoạn và phương pháp end-to-end. Cách tiếp cận này có ưu điểm lớn về chi phí tính toán thấp, khả năng linh hoạt trong việc thay đổi từng thành phần và tốc độ truy vấn nhanh. Tuy nhiên, các mô hình hiện tại vẫn chưa đạt độ chính xác cao do hạn chế trong khả năng hiểu ngữ nghĩa video. Ngược lại, các phương pháp end-to-end đạt độ chính xác rất tốt nhờ cơ chế học biểu diễn video–ngôn ngữ thống nhất. Dù vậy, chúng đòi hỏi chi phí tính toán lớn, tốc độ truy vấn chậm và khó mở rộng khi làm việc với các kho video quy mô lớn.

Dựa trên những phân tích này, đề tài lựa chọn hướng xây dựng mô hình hai giai đoạn cho mục tiêu thứ ba. Trong bước chuẩn bị dữ liệu, mỗi video được trích xuất các khung hình chính (keyframes) với khoảng cách thời gian không vượt quá một giây nhằm đảm bảo không bỏ sót các sự kiện quan trọng. Các khung hình lặp lại hoặc có nội dung giống nhau liên tiếp được loại bỏ bằng mô hình đo độ tương đồng ảnh. Toàn bộ các keyframes sau khi xử lý được lưu trữ vào không gian vector nhằm hỗ trợ quá trình truy vấn nhanh và hiệu quả.

Ở giai đoạn đầu của mô hình, đề tài sử dụng một mô hình ngôn ngữ lớn để phân tách truy vấn thành các sự kiện con theo đúng thứ tự xuất hiện. Chẳng hạn, câu mô tả "Một đầu bếp rửa cà chua và cắt chúng thành các mảnh" sẽ được tách thành hai sự kiện riêng biệt: rửa cà chua và cắt cà chua. Mỗi sự kiện được sử dụng để truy vấn vào không gian vector bằng mô hình đa phương thức nhằm tìm ra top-k keyframes phù hợp nhất. Một video được xem là tiềm năng nếu nó chứa ít nhất một keyframe phù hợp với mỗi sự kiện trong truy vấn. Tập video kèm với các ứng cử viên cho mỗi sự kiện thu được từ quá trình này chính là đầu ra của giai đoạn thứ hai.

Trong giai đoạn thứ hai, đề tài đề xuất cơ chế định vị khoảnh khắc dựa trên quy hoạch động. Với mỗi video tiềm năng, hệ thống duyệt qua các keyframes theo thứ tự thời

gian. Với một keyframe thuộc sự kiện thứ j, mô hình tìm keyframe gần nhất thuộc sự kiện thứ j − 1. Khi xác định được keyframe tương ứng với sự kiện cuối, hệ thống thu được một bộ kết quả hoàn chỉnh thông qua giải pháp trên. Các kết quả này được xếp hạng dựa trên tổng điểm tương đồng giữa các sự kiện và các keyframe liên quan, do mô hình đa phương thức cung cấp.

Đề tài cũng xác định các tiêu chí đánh giá phù hợp cho hai nhiệm vụ cốt lõi. Độ đo Recall@K được sử dụng để đo đạt truy hồi video. Tiêu chí Temporal IoU threshold được áp dụng để đánh giá mức độ trùng khớp giữa đoạn thời gian dự đoán và đoạn thời gian được gán nhãn. Query response time để đánh giá thời gian xử lý truy vấn.

KẾT QUẢ MONG ĐỢI

Từ việc triển khai nội dung và phương pháp đã nêu, đề tài dự kiến xây dựng được một hệ thống Video Corpus Moment Retrieval theo hướng hai giai đoạn với mức độ hoàn chỉnh đủ để đánh giá thử nghiệm. Kết quả đầu tiên là một kho dữ liệu keyframe cho toàn bộ tập video, trong đó mỗi khung hình đã được lọc trùng và biểu diễn bằng véc-tơ đa phương thức, phục vụ cho quá trình truy vấn nhanh và tiết kiệm tài nguyên hơn so với xử lý video thô.

Ở giai đoạn truy hồi video, cho phép người dùng nhập câu truy vấn, hệ thống dự kiến sẽ tách truy vấn thành chuỗi sự kiện bằng mô hình ngôn ngữ lớn và tìm top-k keyframe phù hợp cho từng sự kiện. Từ đó, hệ thống lọc ra danh sách video thỏa điều kiện chứa đầy đủ các sự kiện mô tả trong truy vấn.

Ở giai đoạn định vị, đề tài kỳ vọng triển khai thuật toán quy hoạch động để xác định chuỗi keyframe tuân theo thứ tự sự kiện và suy ra khoảng thời gian bắt đầu – kết thúc

tương ứng trong video. Các kết quả sẽ được xếp hạng dựa trên mức độ tương đồng ngữ nghĩa giữa truy vấn và keyframe.

Cuối cùng, đề tài hướng tới thu được bộ kết quả thực nghiệm gồm Recall@K (cho truy hồi) và IoU hoặc IoU@threshold (cho định vị thời gian), query respone time (thời gian xử lý truy vấn), qua đó đánh giá mức độ chính xác và tốc độ truy vấn của mô hình baseline, làm cơ sở đề xuất các cải tiến ở các nghiên cứu tiếp theo.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Meng Liu, Liqiang Nie, Yunxiao Wang, Qi Tian, Tat-Seng Chua:
A Survey on Video Moment Localization. *ACM Computing Surveys*, 2023

[2] Yawen Zeng, Jieyu Zhang, Haoran Zhang, Xupeng Miao, Zhiwu Lu, Tao Xiang, Yi Yang:
Revisiting Video Corpus Moment Retrieval via Contrastive Learning. SIGIR 2021: 1552–1561.

[3] Souvik Das, Debojyoti Deka, Dheeraj Mekala, Jingbo Shang:
Query-to-Event (Q2E) Decomposition for Zero-Shot Text-to-Video Retrieval. NAACL 2025.

[4]. Mingyu Chen, Wenhan Luo, Lei Jin, Meng Wang:
Event-Aware Video Corpus Moment Retrieval. 2024.

[5] Jinrui Ding, Ye Hao, Jingkuan Song, Lianli Gao, Heng Tao Shen:
Fast Video Moment Retrieval. ICCV 2021: 11651–11660.