ĐẠI HỌC QUỐC GIA TP.HCM
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

Ngày nhận hồ sơ	
(Do CQ quản lý ghi)	

THUYẾT MINH

ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN 2023

A. THÔNG TIN CHUNG

A1. Tên đề tài

- Tên tiếng Việt (IN HOA): NÂNG CAO HIỆU SUẤT TRUY VẤN VIDEO QUA SỰ KẾT HỢP ĐA DẠNG PHƯƠNG PHÁP VÀ CHIẾN LƯỢC TÌM KIẾM THEO THỜI GIAN
- Tên tiếng Anh (IN HOA): ENHANCING VIDEO RETRIEVAL PERFORMANCE THROUGH THE COMBINATION OF DIVERSE METHODS AND TEMPORAL SEARCH STRATEGIES

A2. Thời gian thực hiện

6 tháng (kể từ khi được duyệt).

A3. Tổng kinh phí

Tổng kinh phí: 6 triệu đồng, gồm:

- Kinh phí từ Trường Đại học Công nghệ Thông tin: 6 triệu đồng

A4. Chủ nhiệm

Họ và tên: Trần Gia Bảo

Ngày, tháng, năm sinh: 27/03/2004 Giới tính (Nam/Nữ): Nam

Số CCCD: 056204001556 ; Ngày cấp: 28/04/2021 ; Nơi cấp: Cục trưởng Cục Cảnh sát quản lý hành chính về trật tự xã hội

Mã số sinh viên: 22520121

Số điện thoại liên lạc: 0767143124

Khoa: Khoa học Máy Tính

Số tài khoản: 9767143124 Ngân hàng: Vietcombank

A5. Thành viên đề tài (kể cả CNĐT)

TT	Họ tên	MSSV	Khoa
1	Trần Gia Bảo	22520121	Khoa học Máy tính
2	Trần Nhật Khoa	22520691	Khoa học Máy tính

B. MÔ TẢ NGHIÊN CỨU

B1. Giới thiệu về đề tài

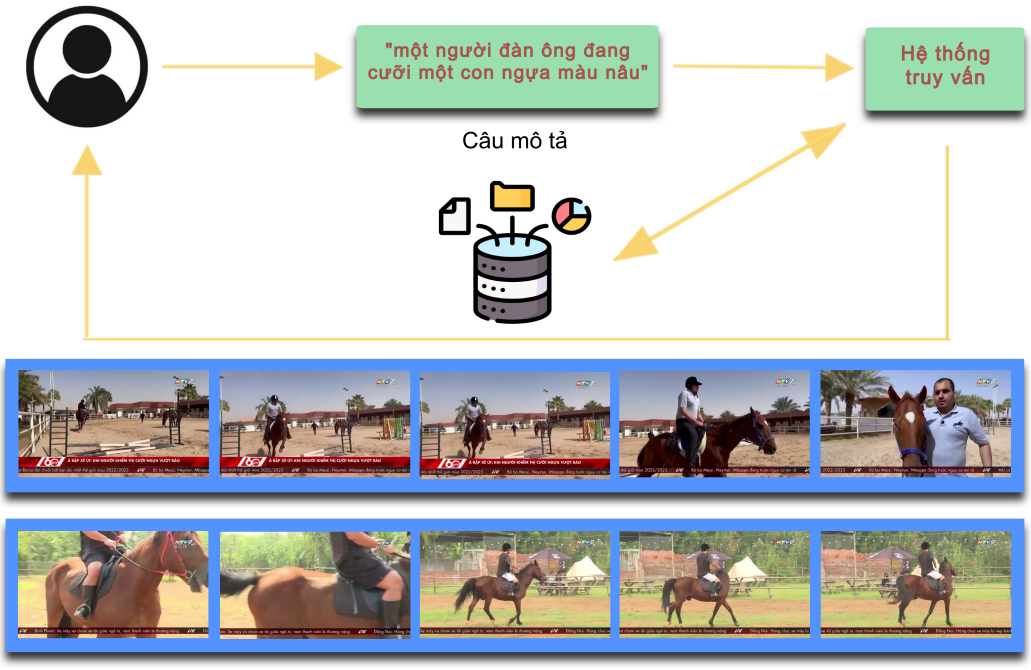
Giới thiệu bài toán

Trong thời đại công nghệ phát triển như hiện nay, sự bùng nổ của dữ liệu video đã trở thành một hiện tượng chưa từng có, thay đổi mọi khía cạnh của cuộc sống từ giao tiếp, giải trí, đến làm việc,... . Từ đó nhu cầu phát triển một hệ thống tìm kiếm video dựa vào câu truy vấn (*text-video retrieval* - TVR) một cách hiệu quả đang thu hút sự chú ý của cộng đồng nghiên cứu gần đây. Đầu vào và đầu ra của một hệ thống truy vấn video sẽ là như sau (xem hình 1):

- Đầu vào bao gồm:
  - Câu truy vấn dưới dạng văn bản.
  - Một tập hợp các video.
- Đầu ra: Một đoạn video ngắn thuộc một trong các video trong tập hợp các video cho trước.

Với kho dữ liệu đầu vào chỉ chứa các video, chúng tôi cần xây dựng một hệ thống cho phép tìm kiếm các đoạn video dựa trên một loạt các mô tả về các thời điểm khác nhau của đoạn video đó, thỏa mãn về nội dung, thông tin về các đối tượng, âm thanh, các đoạn chữ, ... được định nghĩa bởi người dùng. Kết quả của bài toán này có thể được ứng dụng rộng rãi trong nhiều lĩnh vực như an ninh, y tế, giáo dục, giải trí, ... có thể kể đến như sau:

- Tìm kiếm video an ninh: Người dùng có thể tìm kiếm các đoạn video có chứa các hình ảnh, âm thanh, hành động, ... mô tả về đối tượng liên quan đến một sự kiện, một vụ án.
- Tìm kiếm nội dung giải trí: Người dùng muốn tìm kiếm dựa trên một nội dung cụ thể (tin tức/phóng sự, thể thao, phim ảnh/nhạc kịch ...).



Hình 1: Hình minh họa một hệ thống truy vấn video bằng câu mô tả. Hệ thống sẽ nhận đầu vào là câu mô tả, kết quả trả về từ hệ thống truy vấn là danh sách gồm các [videoId, frameID] được sắp xếp dựa trên độ tương đồng cosine. Màn hình hiển thị kết quả trả về là các hình bên dưới.

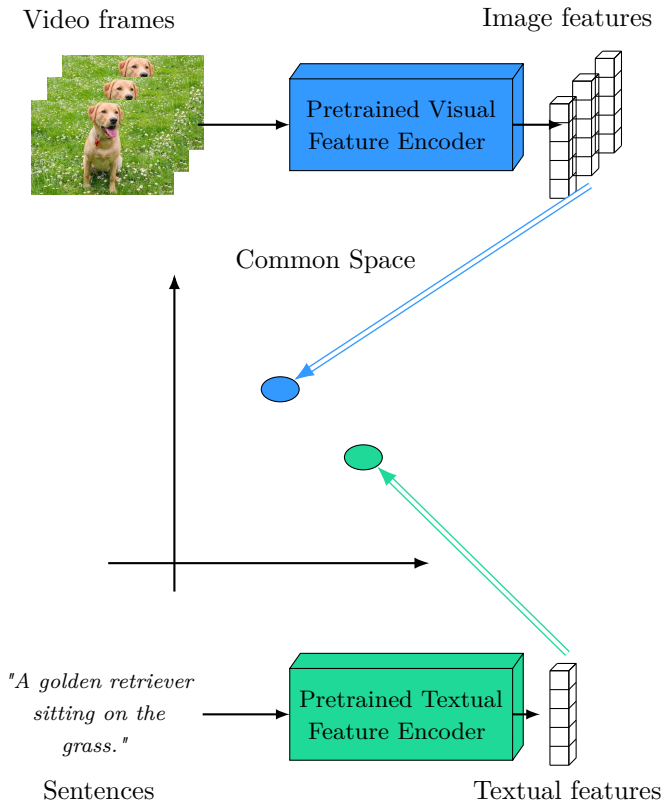
Tổng quan tình hình nghiên cứu

Hiện nay, có nhiều cuộc thi và hội nghị lớn về truy vấn thông tin video, như Lifelog Search Challenge (LSC) tại hội nghị ACM International Conference on Multimedia Retrieval (ACM ICMR) và Video Browser Showdown (VBS) tại hội nghị International Conference on Multimedia Modeling (MMM). Trong những sự kiện này, các đội thi đối mặt với thách thức xây dựng hệ thống tìm kiếm video có khả năng đạt được tốc độ và độ chính xác cao nhất.

Đội thi Vibro[2] sử dụng mô hình CLIP[5] (một mô hình image-text-embedding với mục tiêu nhúng vector biểu diễn của ảnh và vector biểu diễn cho văn bản vào trong cùng một không gian - xem hình 2) ViT-L. Hướng tiếp cận này tìm kiếm thông tin bằng cách đưa đặc trưng thị giác và văn bản về cùng một không gian để tìm kiếm bằng vector, cho thấy sự hiệu quả khi tìm kiếm các tác vụ liên quan đến ảnh. Ngoài ra, đội Vibro còn áp dụng "2D-sorted map", phương pháp có tên self-organizing-map (SOM) để hiển thị kết quả.

Đội VideoCLIP[4] cũng sử dụng mô hình CLIP để tìm kiếm theo câu truy vấn văn bản, dùng đặc trưng hình ảnh để truy vấn hình ảnh tương tự, và dùng Google Vision API để tìm kiếm OCR. Ngoài ra, hệ thống của họ còn hỗ trợ tìm kiếm theo màu chủ đạo (12 màu chủ đạo trong hệ thống của họ).

Hệ thống của đội thi VERGE[1] hỗ trợ việc tìm kiếm tương đồng (visual similarity search) bằng việc áp dụng cả hai phương pháp đầu tiên sử dụng một vector 1024 chiều từ một mô hình GoogleNet (đã được đội họ finetune) và đánh chỉ mục bằng phương pháp IVFADC - cho việc tìm kiếm tương đồng; phương pháp thứ hai được lấy cảm hứng từ phương pháp của Lin và đồng nghiệp [3] (một phương pháp hasing trên không gian Hamming (chỉ chứa giá trị 0 và 1) nhằm nhúng được biểu diễn của ảnh và biểu diễn của text vào cùng không gian Hamming nêu trên) - với đầu vào là vector đặc trưng được rút trích tại tầng cuối (fc7) của VGG16; ngoài ra đội VERGE còn sử dụng kNN để vừa tăng tốc tìm kiếm trên mã hash từ bước trước.



Hình 2: Hình minh họa cách hoạt động của các thuật toán co-embedding. Ở phần đặc trưng ảnh một mô hình CNN hoặc Vision Transformer (thông thường đã được huấn luyện sẵn) nhận đầu vào là video/ảnh, và đầu ra sẽ đặc trưng biểu diễn cho video/ảnh đó. Tương tự đối với phần đặc trưng văn bản.

Mục tiêu đề tài

Vì vậy với mục tiêu xây dựng một hệ thống truy vấn video vừa đảm bảo độ chính xác và hiệu quả về mặt tài nguyên, chúng tôi dựa trên những mô hình xuất sắc từ các giai đoạn trước đó [1, 2, 4]. Ngoài ra, việc thiết kế giao diện cũng được chú trọng để đảm bảo tính đơn giản và dễ sử dụng, đồng thời giúp người dùng tìm kiếm thông tin một cách nhanh chóng và chính xác. Vì vậy, trong khuôn khổ này của đề tài nghiên cứu chúng tôi sẽ thực hiện các nội dung như sau:

- Tìm hiểu tổng quan về bài toán truy xuất video.
- Đánh giá các phương pháp.
- Xây dựng hệ thống truy vấn.

B2. Mục tiêu, nội dung, kế hoạch nghiên cứu

B2.1 Mục tiêu

Chúng tôi đề ra mục tiêu của nghiên cứu này gồm:

- Tìm hiểu tổng quan về bài toán truy xuất video.
- Đánh giá một số phương pháp của các nghiên cứu trước, đồng thời lựa chọn những phương pháp cho kết quả tốt để tích hợp vào hệ thống.
- Xây dựng hệ thống dựa trên các phương pháp đã được chọn, đồng thời tiến hành tinh chỉnh, tối ưu để hệ thống có khả năng truy xuất nhanh hơn, hiệu quả hơn, thân thiện và dễ dàng sử dụng cho người dùng.

B2.2 Nội dung và phương pháp nghiên cứu

Nội dung 1: Tìm hiểu tổng quan về bài toán truy xuất video.

Phương pháp thực hiện:

- Tìm hiểu tổng quan về bài toán truy vấn video;
- Tìm hiểu các hướng tiếp cận tiên tiến gần đây để giải quyết bài toán;
- Tìm hiểu các hệ thống đạt giải thưởng cao trong các cuộc thi về bài toán này như Video Browser Showdown, Lifelog Search Challenge.

Kết quả dự kiến:

- Báo cáo tổng quan về bài toán truy vấn sự kiện từ video.
- Báo cáo tổng quan về những phương pháp giúp việc tìm kiếm trở nên hiệu quả hơn.

Nội dung 2: Đánh giá các phương pháp

Phương pháp thực hiện:

- Thực nghiệm các phương pháp trên các bộ dữ liệu tương ứng cho từng phương pháp.

Kết quả dự kiến:

- Tài liệu về cách đánh giá các phương pháp.
- Bảng kết quả của các phương pháp.

Nội dung 3: Xây dựng hệ thống truy vấn

Phương pháp nghiên cứu:

- Sử dụng các công cụ để xây dựng hệ thống truy vấn bao gồm các chức năng trên.
- Tối ưu hóa giao diện, tốc độ xử lý để giúp việc tìm kiếm hiệu quả hơn.

Kết quả dự kiến:

- Một hệ thống web truy vấn video hoàn chỉnh, hiệu quả, nhanh chóng, thân thiện với người dùng.

B3. Kết quả dự kiến

Các kết quả dự kiến của đề tài này là:

- Tài liệu báo cáo tổng quan về bài toán truy vấn sự kiện từ video
- Tài liệu phân tích kết quả đánh giá của các phương pháp
- Xây dựng một hệ thống web truy vấn hoàn chỉnh.

B4. Tài liệu tham khảo

[1] Stelios Andreadis, Anastasia Moumtzidou, Damianos Galanopoulos, Nick Pantelidis, Konstantinos Apostolidis, Despoina Touska, Konstantinos Gkountakos, Maria Pegia, Ilias Gialampoukidis, Stefanos Vrochidis, Vasileios Mezaris, and Ioannis Kompatsiaris. "VERGE in VBS 2022". In: *MultiMedia Modeling*. Springer International Publishing, 2022, pp. 530–536. DOI: 10.1007/978-3-030-98355-0_50. URL: https://doi.org/10.1007/978-3-030-98355-0_50.

[2] Nico Hezel, Konstantin Schall, Klaus Jung, and Kai Uwe Barthel. "Efficient Search and Browsing of Large-Scale Video Collections with Vibro". In: *MultiMedia Modeling*. Springer International Publishing, 2022, pp. 487–492. DOI: 10.1007/978-3-030-98355-0_43. URL: https://doi.org/10.1007/978-3-030-98355-0_43.

[3] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. "Semantics-preserving hashing for cross-view retrieval". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3864–3872.

[4] Thao-Nhu Nguyen, Bunyarit Puangthamawathanakun, Annalina Caputo, Graham Healy, Binh T. Nguyen, Chonlameth Arpnikanondt, and Cathal Gurrin. "VideoCLIP: An Interactive CLIP-based Video Retrieval System at VBS2023". In: *MultiMedia Modeling*. Springer International Publishing, 2023, pp. 671–677. DOI: 10.1007/978-3-031-27077-2_57. URL: https://doi.org/10.1007/978-3-031-27077-2_57.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

Ngày . . . tháng . . . năm 2023	*Ngày . . . tháng . . . năm 2023*
Giảng viên hướng dẫn	**Chủ nhiệm đề tài**
(Ký và ghi rõ họ tên)	(Ký và ghi rõ họ tên)