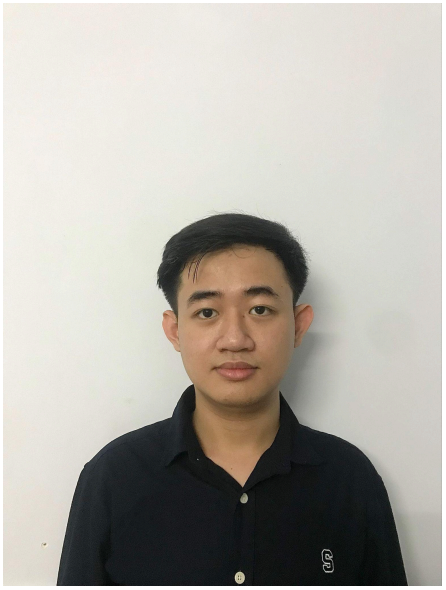
THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/NaNFNEbIBsw*
- Link slides (dạng .pdf đặt trên Github của nhóm): *https://github.com/tai040102/-CS2205.FEB2025/blob/main/T%C3%A0i%20Ph%E1%BA%A1m%20T%E1%BA%A5n_CS2205.FEB2025.DeCuong.FinalReport.Template.Slide.pdf*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

- Họ và Tên: Phạm Tấn Tài
- MSSV: 240101069



- Lớp: CS2205.CH190
- Tự đánh giá (điểm tổng kết môn):9/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 4
- Link Github: https://github.com/tai040102/-CS2205.FEB2025

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

TĂNG CƯỜNG KHẢ NĂNG PHÁT HIỆN TƯƠNG TÁC NGƯỜI VẬT THÔNG QUA ĐẶC TRƯNG NGỮ NGHĨA VÀ NON QUERY-KEY-VALUE TRONG CÁC MÔ HÌNH TRANSFORMER.

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

ENHANCING HUMAN-OBJECT INTERACTION DETECTION VIA SEMANTIC FEATURES AND NON QUERY-KEY-VALUE MECHANISMS IN TRANSFORMER MODELS.

TÓM TẮT *(Tối đa 400 từ)*

Bài toán phát hiện tương tác người–vật là một hướng nghiên cứu trong lĩnh vực thị giác máy tính, nhằm xác định không chỉ vị trí của người và vật thể trong ảnh, mà còn hiểu được các mối quan hệ tương tác giữa họ. Mặc dù đã có nhiều tiến bộ trong lĩnh vực này, nhưng những thách thức như tính đa dạng, phức tạp của các tình huống tương tác, và việc tận dụng hiệu quả thông tin ngữ nghĩa trong mô hình vẫn còn là vấn đề mở.

Trong thời gian gần đây, các mô hình sử dụng kiến trúc Transformer như SOV-STG[1], PViC[2] đã cho thấy hiệu quả cao trong bài toán HOI, đặc biệt khi được xây dựng trên các kiến trúc mạnh mẽ như Swin Transformer[5]. Tuy nhiên, các mô hình hiện tại vẫn còn hạn chế trong việc khai thác triệt để đặc trưng ngữ nghĩa, và quá phụ thuộc vào cơ chế attention truyền thống với bộ ba Query-Key-Value (QKV).

Đề tài này tập trung khảo sát các hướng tiếp cận hiện đại trong HOI, với trọng tâm là các kỹ thuật trích xuất và khai thác đặc trưng ngữ nghĩa từ hình ảnh. Trên cơ sở đó, đề xuất một phương pháp cải tiến mô hình Transformer nhằm tăng cường khả năng phát hiện tương tác người–vật thông qua việc tích hợp thêm thông tin ngữ nghĩa vào bộ giải mã, kết hợp với việc nghiên cứu cơ chế phi QKV nhằm tạo ra biểu diễn tương tác hiệu quả hơn.

Việc đánh giá hiệu quả của mô hình sẽ được thực hiện trên các bộ dữ liệu tiêu chuẩn như V-COCO[3] và HICO-DET[4]. Cuối cùng, một chương trình ứng dụng minh họa sẽ được phát triển nhằm trực quan hóa kết quả mô hình và kiểm nghiệm tính khả thi trong thực tế.

Kết quả mong đợi của đề tài bao gồm: báo cáo nghiên cứu chi tiết, mã nguồn mô hình cùng tài liệu hướng dẫn triển khai, và một ứng dụng minh họa kết quả phát hiện tương tác người–vật.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Bài toán phát hiện tương tác người vật là một trong những bài toán trong lĩnh vực thị giác máy tính. Nó tập trung vào việc nhận diện và hiểu sự tương tác giữa con người và vật thể trong một hình ảnh. Bài toán này không chỉ đơn thuần là phát hiện và định vị người và vật thể mà còn đòi hỏi hiểu được các mối quan hệ và tương tác giữa họ, chẳng hạn như việc một người cầm một đối tượng, lái một chiếc xe, hoặc tiếp xúc vật thể trong các tình huống thực tế. Phát hiện tương tác người-vật có ứng dụng rộng rãi trong nhiều lĩnh vực, bao gồm nhận dạng hành vi con người, giám sát an ninh và nhiều ứng dụng thú vị khác. Bài toán HOI(Human Object Interaction) đặt ra những thách thức đáng kể do sự đa dạng và phức tạp của các tình huống tương tác giữa người và vật thể. Các tình huống này rất đa dạng và có thể xuất hiện dưới nhiều hình thức khác nhau:

- Nhiều người với nhiều tương tác đa dạng: Ví dụ, trong một bữa tiệc, nhiều người có thể đang nói chuyện, cười đùa, và cùng nhau thưởng thức thức ăn. Điều này đòi hỏi mô hình phải xác định được các loại tương tác đa dạng và lúc này các tương tác có thể bị chồng lấp lên nhau.
- Một người tương tác với nhiều vật thể cùng lúc: Một ví dụ cho trường hợp này là khi người dùng ngồi trên một chiếc ghế và đang sử dụng máy tính. Mô hình cần xác định được rằng người đó không chỉ ngồi trên ghế mà còn đang tương tác với máy tính.
- Nhiều người có cùng một tương tác với vật thể: Ví dụ, khi một nhóm bạn đang

ném và bắt một quả bóng, cả nhóm chia sẻ cùng một tương tác với quả bóng. Điều này đòi hỏi mô hình phải hiểu được tương tác chung của nhiều người với vật thể (quả bóng) này.

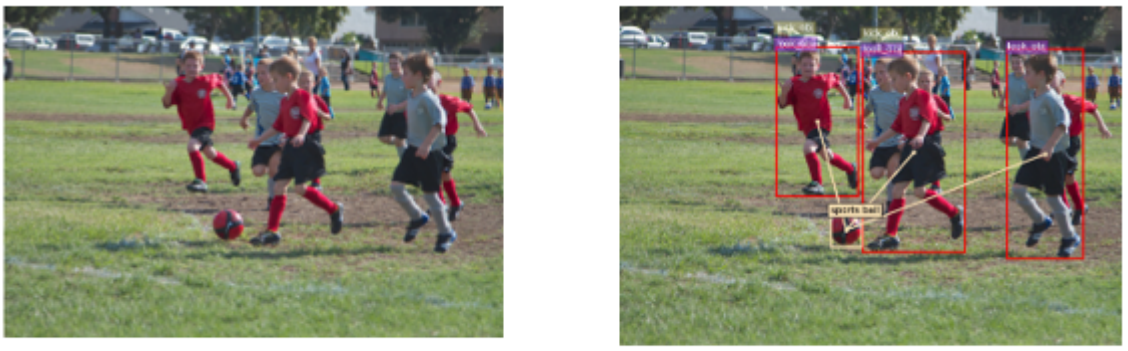
Trong thực tế, có rất nhiều tình huống phức tạp và đa dạng, làm cho việc phát hiện và hiểu các mối tương tác giữa người và vật thể trở thành một thách thức lớn trong lĩnh vực Thị giác máy tính.

Trong các mô hình giải quyết bài toán này mà sử dụng kiến trúc transformer nổi tiếng có thể kể đến là SOV-STG[1],PViC[2] đây là các mô hình được phát triển lên từ Swin Transformer[5], các mô hình này đã cho thấy rằng nó có thể đạt được hiệu suất rất tốt. Nhưng hiện tại trong các kiến trúc của các mô hình ở trên thì decoder vẫn chưa khai thác triệt để được thông tin của các đặc trưng ngữ nghĩa cũng như việc tạo affinity matrix vẫn chỉ phụ thuộc vào bộ ba QKV. Thấy được tiềm năng hiện tại của bài toán nên chúng tôi quyết định lựa chọn bài toán này cho việc tìm hiểu và nghiên cứu.

Phát biểu bài toán:

Đầu vào: Một hình ảnh trong đó một (hoặc nhiều) người đang tham gia tương tác với một (hoặc nhiều) vật thể.

Đầu ra: Phát hiện và nhận diện tập hợp chứa các bộ ba: người, vật thể và các tương tác giữa họ, được biểu diễn dưới dạng {human, object, interaction}.



(a) Đầu vào (b) Đầu ra

Hình 1.1 Đầu vào và đầu ra của bài toán

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

Nâng cao hiệu quả phát hiện tương tác người – vật bằng cách khai thác đặc trưng ngữ

nghĩa.

Nâng cao hiệu quả phát hiện tương tác người – vật bằng cách sử dụng các cơ chế non QKV.

Đề xuất kiến trúc HOI Transformer cải tiến giúp tăng độ chính xác và khả năng tổng quát trên các bộ dữ liệu chuẩn như VCOCO[3] và HICO-DET[4].

NỘI DUNG VÀ PHƯƠNG PHÁP

Nội dung của đề tài tập trung vào việc khảo sát và đề xuất cải tiến trong bài toán phát hiện tương tác người – vật dựa trên các đặc trưng ngữ nghĩa và cơ chế non Query-Key-Value trong mô hình Transformer.

Trước tiên, đề tài tiến hành nghiên cứu tổng quan các hướng tiếp cận hiện đại (state-of-the-art) trong HOI, đặc biệt là các phương pháp khai thác đặc trưng ngữ nghĩa từ hình ảnh và các phương pháp sử dụng non QKV như [6],[7] . Từ đó, lựa chọn và cài đặt lại một số mô hình tiêu biểu có tiềm năng cải thiện hiệu suất. Đặc biệt, PViC[2] sẽ được phân tích chuyên sâu nhằm đề xuất một phương pháp cải tiến khả năng tận dụng đặc trưng ngữ nghĩa ở cả bộ mã hóa và giải mã .

Đề tài cũng tìm hiểu về các tập dữ liệu chuẩn như V-COCO[3] và HICO-DET[4] để sử dụng trong huấn luyện và đánh giá mô hình. Các mô hình sau khi thiết kế và cải tiến sẽ được thực nghiệm trên các tập dữ liệu này, tiến hành tổng hợp kết quả và so sánh với các phương pháp trước đó nhằm đánh giá mức độ cải thiện.

Cuối cùng, một ứng dụng minh họa sẽ được xây dựng để trình bày khả năng áp dụng thực tế của mô hình đề xuất.

KẾT QUẢ MONG ĐỢI

Bản báo cáo chi tiết cung cấp một cái nhìn toàn diện về quá trình tìm hiểu và khảo sát trong bài toán Phát Hiện Tương Tác Người-Vật của chúng tôi. Báo cáo này sẽ không chỉ phân tích mà còn đề xuất phương pháp để cải thiện hiệu suất trong HOI.

Source code và hướng dẫn cài đặt chi tiết về mô hình đề xuất.

Xây dựng chương trình ứng dụng minh họa để trực quan hóa kết quả nghiên cứu.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1].Chen, Junwen, Yingcheng Wang, and Keiji Yanai. "Focusing on what to Decode and what to Train: SOV Decoding with Specific Target Guided DeNoising and Vision Language Advisor." 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2025.

[2]. Zhang, Frederic Z., et al. "Exploring predicate visual context in detecting of human-object interactions." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

[3]. Gupta, Saurabh, and Jitendra Malik. "Visual semantic role labeling." arXiv preprint arXiv:1505.04474 (2015).

[4]. Chao, Yu-Wei, et al. "Learning to detect human-object interactions." 2018 ieee winter conference on applications of computer vision (wacv). IEEE, 2018.

[5]. Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[6]. Duo Li, Jie Hu, Changhu Wang, Xiangtai Li, Qi She, Lei Zhu, Tong Zhang, and Qifeng Chen. Involution: Inverting the inherence of convolution for visual recognition. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pages 12321–12330. Computer Vision Foundation / IEEE, 2021.

[7]. Arar, Moab, Ariel Shamir, and Amit H. Bermano. "Learned queries for efficient local attention." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.