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

Môn học: CS2205 - PHƯƠNG PHÁP LUẬN NCKH

Lớp: CS2205.FEB2025

GV: PGS.TS. Lê Đình Duy

Trường ĐH Công Nghệ Thông Tin, ĐHQG-HCM

Tóm tắt

- Link Github của nhóm: https://github.com/tai040102/-CS2205.FEB2025
- Link YouTube video: https://youtu.be/NaNFNEbIBsw
- Phạm Tấn Tài - 240101069

Giới thiệu

- Bài toán HOI (Human-Object Interaction) là một nhánh quan trọng trong thị giác máy tính, nhằm nhận diện và hiểu tương tác giữa con người và vật thể trong ảnh.
- Ứng dụng: Giám sát an ninh và phân tích hành vi.
- Thách thức:
  - Nhiều người cùng tương tác với một vật thể.
  - Một người tương tác đồng thời nhiều vật thể.
  - Nhiều người với nhiều tương tác đa dạng.

Giới thiệu

Phát biểu bài toán:

- Đầu vào: Bức ảnh chứa một (nhiều) người tương tác với một (nhiều) đối tượng.
- Đầu ra: Tập các bộ ba {H, O, I} gồm:Hộp giới hạn cho người.Hộp giới hạn và nhãn cho đối tượng.Loại tương tác giữa người và đối tượng đó.

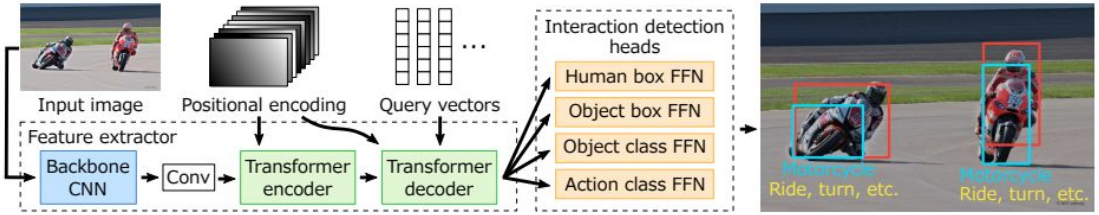
(a) Đầu vào

(b) Đầu ra

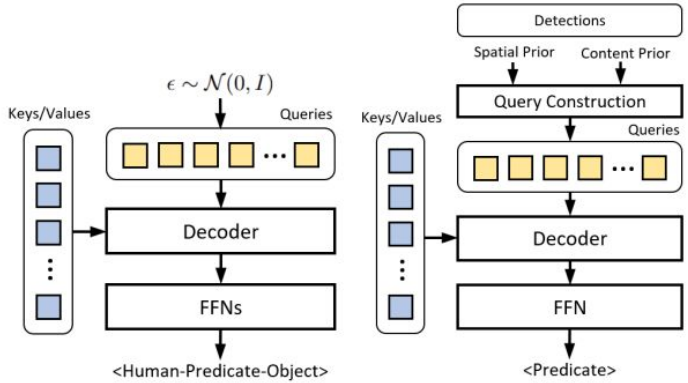
Giới thiệu

Các mô hình hiện đại như SOV-STG và PViC dựa trên Swin Transformer đạt kết quả tốt, nhưng:

- Phụ thuộc vào cơ chế QKV truyền thống.
- Chưa khai thác hiệu quả đặc trưng ngữ nghĩa ở decoder.



Kiến trúc Transformer cho bài toán HOI



Kiến trúc Decoder của PViC

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 và HICO-DET.

Nội dung và Phương pháp

- Khảo sát các phương pháp HOI hiện đại, đặc biệt tập trung vào những mô hình sử dụng đặc trưng ngữ nghĩa và cơ chế Non-QKV như QLV và LKV.
- Phân tích sâu kiến trúc PViC – một mô hình HOI hiệu quả dựa trên Swin Transformer – nhằm nhận diện các hạn chế và tiềm năng cải tiến.
- Đề xuất kiến trúc HOI Transformer mới, tích hợp semantic features ở decoder, đồng thời thay thế cơ chế QKV truyền thống bằng Non-QKV.

Nội dung và Phương pháp

- Huấn luyện và đánh giá mô hình trên các tập dữ liệu chuẩn như V-COCO và HICO-DET, so sánh với các phương pháp SOTA hiện tại.
- Xây dựng ứng dụng minh họa, thể hiện khả năng ứng dụng mô hình vào các bài toán thực tế như giám sát hoặc nhận diện hành vi.

Kết quả dự kiến

- Bản báo cáo chi tiết.
- Source code và hướng dẫn cài đặt chi tiết.
- Xây dựng chương trình ứng dụng minh họa.

Tài liệu tham khảo

[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]. Tamura, Masato, Hiroki Ohashi, and Tomoaki Yoshinaga. "Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

Tài liệu tham khảo

[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.