NGHIÊN CỨU VÀ PHÁT TRIỂN MÔ HÌNH HỌC SÂU CHO BÀI TOÁN PHÁT HIỆN VÀ NHẬN DIỆN HÀNH ĐỘNG CỦA CON NGƯỜI.

Nguyễn Đặng Đức Mạnh - 22520847

Tóm tắt

- Lớp: **CS519.P11**
- Link Github của nhóm: https://github.com/Hope1337/CS519.P11
- Link YouTube video: https://www.youtube.com/watch?v=LKdcdXxWJGA
- Ảnh + Họ và Tên của các thành viên : Nguyễn Đặng Đức Mạnh

Giới thiệu

- •Trong lĩnh vực thị giác máy tính, HADR (Human Action Detection and Recognition) là tác vụ yêu cầu mô hình phát hiện (bounding box) và nhận diện (phân loại) hành động của con người từ một clip ngắn cho trước.

Giới thiệu

Tại sao lại cần bounding box :

- ***Label*** : Có ích trong các tác vụ phân loại video, nằm ở mức tổng quát về nội dung của video.

- ***Starting + ending point*** : có ích trong bài toán chỉ yêu cầu truy vấn xem hành động cho trước xảy ra ở đâu trong video.

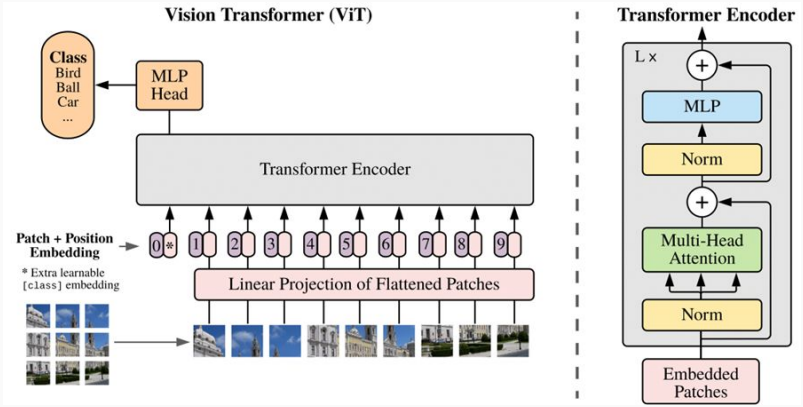
- ***Bounding box*** : có ý nghĩa trong các tác vụ truy vấn cụ thể hơn (ví dụ như người A đang thực hiện hành động B), tận dụng tốt các thông tin không gian trong từng frame.

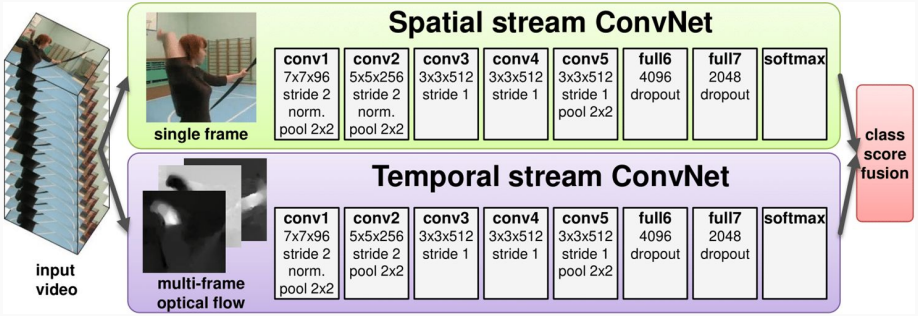
Mục tiêu

- Tìm hiểu tổng quan về bài toán phát hiện và nhận diện hành động con người (HADR).
- Phân tích, đánh giá hiệu suất của ứng dụng mô hình học sâu thời gian thực cho bài toán HADR.
- Đề xuất mô hình học sâu hai luồng (Two-Stream Network) dựa vào kiến trúc mạng gồm ba thành phần: Backbone (trích xuất đặc trưng), Neck (tăng cường đặc trưng đa tỉ lệ) và Head (Lớp tích chập với nhiệm vụ phát hiện và nhận diện) nhằm nâng cao độ chính xác nhưng vẫn giữ được tốc độ xử lý theo thời gian của hệ thống phát hiện và nhận diện hành động.

Nội dung và Phương pháp

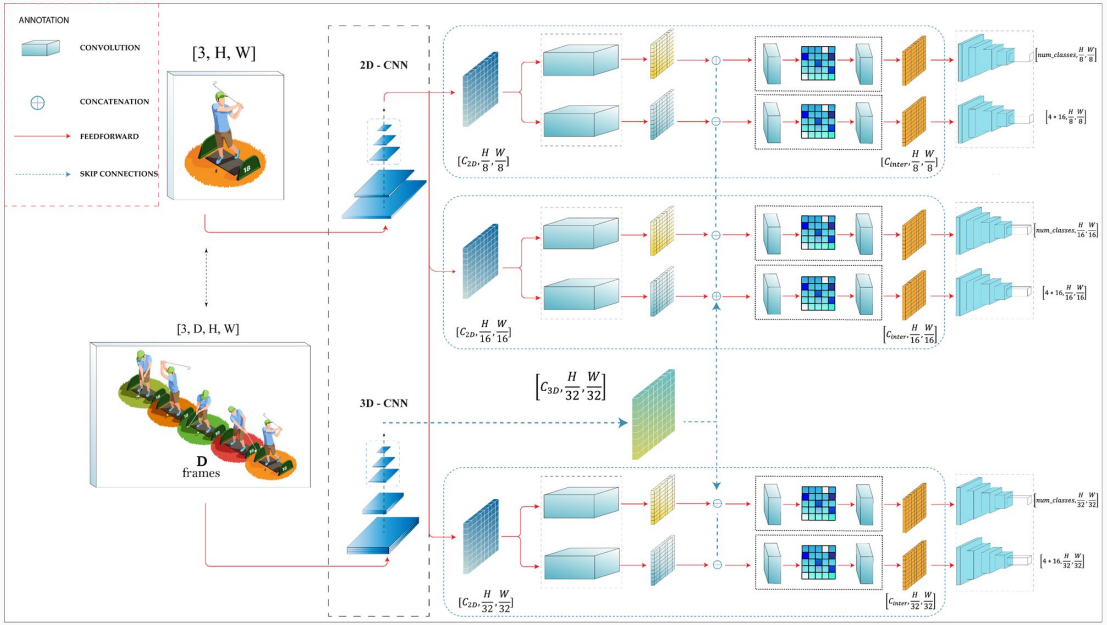
Nội dung 1: Tìm hiểu tổng quan về bài toán phát hiện và nhận diện hành động





- Khảo sát, tìm hiểu tổng quan bài toán thông qua các nghiên cứu mới nhất công bố trên các kỉ yếu hội thảo khoa học chuyên ngành và các tạp chí có uy tín.
- Khảo sát, tìm hiểu một số hướng tiếp cận tiên tiến hiện nay cho bài toán phát hiện và nhận diện hành động dựa trên học sâu.

Nội dung và Phương pháp



Nội dung 2: Đề xuất mô hình hai luồng (Two-Stream Network) :

Đề xuất mô hình mạng học sâu hai luồng (Two-Stream Network) với ba thành phần Backbone, Neck, Head. Backbone đảm nhận nhiệm vụ trích xuất đặc trưng, gồm hai luồng xử lí: thông tin về chuyển động theo thời gian (kiến trúc 3D CNN) và luồng thông tin về không gian (2D CNN). Neck áp dụng kiến trúc kim tự tháp (Feature Pyramid Network) để thực hiện nhiệm vụ kết hợp những đặc trưng từ các lớp tích chập khác nhau để tăng cường ngữ nghĩa. Head đóng vai trò nhận thông tin từ Neck để thực hiện phát hiện và nhận diện hành động

Nội dung và Phương pháp

Nội dung 3: Thiết kế thực nghiệm và phân tích đánh giá

- Huấn luyện và đánh giá mô hình được đề xuất cải tiến dựa trên so sánh với các benchmark , sử dụng các metric và các tập dữ liệu phát hiện và nhận diện hành động phổ biến.

Bảng 1. Tập Datasets dự kiến được sử dụng thực nghiệm

Name	Num classes	Total Sample	Train	Val	Test
UCF101-24	24	>400K key frame	>300K key frame	None	>100K key frame
AVA v2.2	80	430 videos	235 videos	64 videos	131 videos

UCF101-24 : Được trích từ tập UCF101 và được label cho task HADR.

AVA v2.2 : Được Google Research public cho cộng đồng nghiên cứu cho task HADR.

Kết quả dự kiến

- Một kiến trúc mạng học sâu Two-Streams Network được đề xuất cải tiến và kiểm chứng bằng thực nghiệm, hoạt động tốt trong việc phát hiện và nhận diện hành động con người trong video, hướng đến nâng cao hiệu suất của các ứng dụng thời gian thực.

- Một bài báo gửi tham dự Hội thảo quốc tế chuyên ngành.

Tài liệu tham khảo

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, RealTime Object Detection," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[2] G. Jocher, A. Chaurasia, and J. Qiu. Ultralytics YOLO (Version 8.0.0), 2023 [Computer software]. https://github.com/ultralytics/ultralytics.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv: 2010. 11929 [cs.CV], 2021.

[4] O. Köpüklü, X. Wei, and G. Rigoll "You Only Watch Once: A Unified CNN Architecture for Real-Time Spatiotemporal Action Localization," 2021, https://arxiv.org/abs/1911.06644.

[5] J. Yang, and K. Dai, "YOWOv2: A Stronger yet Efficient Multi-level Detection Framework for Real-time Spatio-temporal Action Detection," 2023, https://arxiv.org/abs/2302.06848.

[6] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression," The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), 2020.

[7] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully Convolutional One-Stage Object Detection," IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[8] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "TOOD: Task-aligned One-stage Object Detection," IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[9] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection," The 34th International Conference on Neural Information Processing Systems (NIPS), 2020.

[10] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 318 – 327, 2020.