


THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
<https://www.youtube.com/watch?v=LKdcdXxWJGA>
- Link slides (dạng .pdf đặt trên Github của nhóm):
<https://github.com/Hope1337/CS519.P11>
- *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*

<ul style="list-style-type: none">● Họ và Tên: Nguyễn Đặng Đức Mạnh● MSSV: 22520847 	<ul style="list-style-type: none">● Lớp: CS519.P11● Tự đánh giá (điểm tổng kết môn): 9.5/10● Số buổi vắng: 1● Số câu hỏi QT cá nhân: 5● Số câu hỏi QT của cả nhóm: 5● Link Github: https://github.com/Hope1337/CS519.P11● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none">○ Tìm hiểu về đề tài, lên ý tưởng○ Viết proposal○ Thiết kế poster○ Thiết kế slide○ Làm video thuyết trình
---	--

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

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

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

RESEARCH AND DEVELOPMENT OF A DEEP LEARNING MODEL FOR THE PROBLEM OF DETECTING AND RECOGNIZING HUMAN ACTIONS

TÓM TẮT (*Tối đa 400 từ*)

Bài toán phát hiện và nhận diện hành động của con người, gọi tắt là HADR (Human Action Detection and Recognition), là tác vụ yêu cầu từ một đoạn clip ngắn cho trước, mô hình cần xác định được người thực hiện hành động đang ở đâu trong khung hình (phát hiện) và phân loại đó là hành động gì (nhận diện). HADR có ý nghĩa lớn trong các tác vụ truy vấn ở mức chi tiết khi đòi hỏi mô hình hiểu đồng thời cả về thông tin về không gian lẫn thời gian.

Các phương pháp hiện đang cho thấy hiệu suất tốt nhất đối với HADR được kế thừa từ lĩnh vực xử lý ngôn ngữ tự nhiên: Vision Transformers (ViT) [3]. Tuy nhiên ViT đòi hỏi một khối lượng tính toán khổng lồ, do đó không thể được ứng dụng rộng rãi cho các tác vụ đòi hỏi tốc độ xử lý thời gian thực. Cũng giống như khi sự quan tâm cho Two-stage detector dần chuyển sang One-stage detector trong Object Detection, một phương pháp khác đã được sinh ra để nâng cấp và cải tiến so với ViT về mặt tốc độ: Two-Streams Network.

Two-Streams Network là một cách thiết kế mạng CNN gồm hai nhánh xử lý: 2D branch – trích xuất đặc trưng không gian (Spatial feature) và 3D branch – trích xuất đặc trưng chuyển động theo thời gian (Temporal feature). Two-Stream Network tỏ ra hiệu quả trong bài toán Action Detection với yêu cầu xử lý theo thời gian thực với

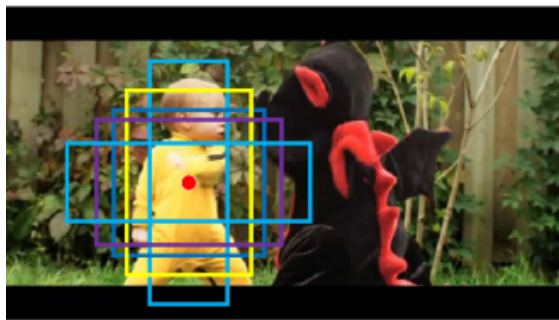
tiêu biểu nhất là mô hình YOWO [4] và YOWOv2 [5].

Tuy nhiên kiến trúc của YOWO [4] và YOWOv2 [5] đều còn khá đơn giản, hơn nữa hiện nay với sự xuất hiện của YOLOv8 và các kỹ thuật mới được đề xuất như Distribution Loss [6], Free Anchor [7], Task Alignment Learning [8], CIoU Loss [9], ... đã chứng minh hiệu quả trong các bài toán thị giác máy tính. Nhận thấy được những tiềm năng ứng dụng các phương pháp cải tiến và kết quả sơ khởi, chúng tôi quyết định chọn “Nghiên cứu và phát triển mô hình học sâu cho bài toán nhận diện và phát hiện hành động” làm đề tài nghiên cứu của mình.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Chúng tôi ứng dụng một số kỹ thuật nổi tiếng trong cộng đồng thị giác máy tính như Anchor-Free model [7], Distribution Loss [9], CIoU Loss [9], TAL [8] và EMA để cải tiến hiệu suất của mô hình Two-Stream Network. Cụ thể, đề tài tập trung một số vấn đề sau:

- Thay đổi cơ chế Anchor box qua Anchor Free: như đã được chứng minh ở [10], các Detector dựa trên cơ chế Anchor Box rất nhạy cảm với hyperameters này (tỉ lệ, kích thước), việc chọn Anchor Box như thế nào cần thông qua các phân tích cho từng tập dữ liệu cụ thể, gây khó khăn cho người thiết kế cũng như quá trình học của mô hình. Cơ chế Anchor Free [7] loại bỏ được công đoạn định nghĩa các Anchor Box mà vẫn giữ được độ chính xác tương đương, giúp mô hình mang tính tổng quát cao hơn cũng như thích nghi với từng bộ dữ liệu khác nhau mà không cần phải qua nhiều bước phân tích tiên lượng.

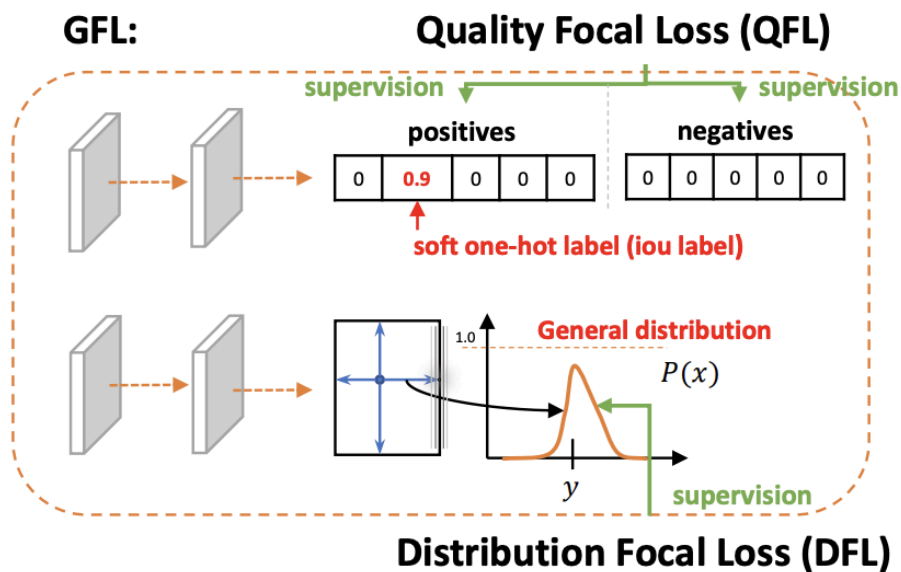


Anchor-based



Anchor-free

- Kết hợp Distribution Loss cho Box Regression: Nhóm tác giả ở [9] lập luận rằng các mô hình hiện nay chỉ đơn thuần dự đoán Bounding Box qua phân phối Dirac Delta hoặc Gaussian, hai phân phối này vẫn còn quá đơn giản và chưa phản ánh được thực sự phân phối của các Bounding Box, do đó các tác giả đề xuất cho mô hình học ra một phân phối tổng quát $P(x)$ và một hàm Distribution Loss cho việc dự đoán phân phối $P(x)$ này.



- Sử dụng CIoU Loss cho Box Regression: Nhóm tác giả ở [6] nhận thấy việc sử dụng MAE hoặc L2 distance cho Box Regression ở giai đoạn train và dùng IoU để đánh giá ở giai đoạn evaluate là không đồng nhất, hơn nữa IoU cũng chưa thực sự phản ánh tốt được việc dự đoán Box tốt hay chưa tốt. Do đó, DIoU, GIoU và CIoU được đề xuất để giải quyết vấn đề này.

- Tăng cường chất lượng gán mẫu Positive và Negative bằng cơ chế Task Alignment Learning: Nghiên cứu ở [8] cho thấy khoảng cách giữa Anchor-Free base và Anchor base là do ta định nghĩa như thế nào là mẫu Positive và Negative, từ đó một chiến thuật tối ưu hơn để gán nhãn cho mẫu được đề xuất.
- Tăng cường tính ổn định của mô hình thông qua cơ chế Exponential Moving Average (EMA): Trong các dữ liệu theo thời gian, việc làm mượt hàm biểu diễn cũng giúp mô hình học tốt hơn. Mặt khác Mini Batch Stochastic Gradient Descent là thuật toán phổ biến dùng để xấp xỉ gradient và tối ưu chi phí tính toán cho mô hình, tuy nhiên việc làm này cũng khiến cho gradient bị nhiễu loạn, cơ chế EMA được áp dụng để hạn chế vấn đề trên.

MỤC TIÊU (*Viết trong vòng 3 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

Mục tiêu nội dung 1 :

- Khảo sát, tìm hiểu tổng quan bài toán nhận diện và phát hiện hành động.
- Phân tích các phương pháp học sâu phổ biến được áp dụng cho bài toán nhận

diện và phát hiện hành động thông qua các nghiên cứu mới nhất.

- Tìm hiểu một số tập dữ liệu nhận diện và phát hiện hành động phổ biến

Sản phẩm dự kiến và chỉ tiêu đánh giá :

- Tài liệu tổng quan về bài toán phát hiện và nhận diện hành động.
- Tài liệu về tập dữ liệu sử dụng cho bài toán.

Phương pháp thực hiện:

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

Mục tiêu nội dung 2 :

- 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

Sản phẩm dự kiến và chỉ tiêu đánh giá

- Kiến trúc mô hình Two-Stream Network được cải tiến, huấn luyện và đánh giá.

Phương pháp thực hiện

- Nghiên cứu lý thuyết, xác định cơ sở lý thuyết cho mô hình đề xuất.

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

Mục tiêu nội dung 3

- 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.

Sản phẩm dự kiến và chỉ tiêu đánh giá

- Hệ thống phát hiện và nhận dạng hoàn chỉnh với các siêu tham số được cập nhật phù hợp.
- Báo cáo tổng kết sơ khởi .

Phương pháp thực hiện

- Thực nghiệm, đánh giá, phân tích kết quả thực nghiệm làm cơ sở kiểm chứng mô hình.

KẾT QUẢ MONG ĐỢI

- 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.