# **THÔNG TIN CHUNG CỦA BÁO CÁO**

* Link YouTube video của báo cáo (tối đa 5 phút):   
  *(ví dụ: https://www.youtube.com/watch?v=AWq7uw-36Ng)*
* Link slides (dạng .pdf đặt trên Github):   
  *(ví dụ: https://github.com/mynameuit/CS2205.APR2023/TenDeTai.pdf)*
* *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*

|  |  |
| --- | --- |
| * Họ và Tên: Nguyễn Tấn Phúc * MSSV: 220101035   A person standing in front of a wall of rugs  Description automatically generated | * Lớp: CS2205.APR2023 * Tự đánh giá (điểm tổng kết môn): 7.5/10 * Số buổi vắng: 0 * Số câu hỏi QT cá nhân: 2 * Số câu hỏi QT của cả nhóm: 5 * Link Github: * Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:   + Lên ý tưởng: Nguyễn Tấn Phúc   + Triển khai: Nguyễn Tấn Phúc   + Làm video YouTube: Nguyễn Tấn Phúc |

# **ĐỀ CƯƠNG NGHIÊN CỨU**

|  |
| --- |
| **TÊN ĐỀ TÀI**  YOLACT: PHÂN VÙNG THỰC THỂ THỜI GIAN THỰC |
| **TÊN ĐỀ TÀI TIẾNG ANH**  YOLACT: REAL-TIME INSTANCE SEGMENTATION |
| **TÓM TẮT**  Phân vùng thực thể là một bài toán thách thức trong lĩnh vực thị giác máy tính nói riêng cũng như ngành khoa học máy tính nói chung, với mục tiêu xác định và phân vùng từng cá thể riêng biệt trong một bức ảnh đầu vào. Trong bài báo này, chúng tôi đề xuất một giải pháp mới, đơn giản gọi là *YOLACT* (**Y**ou **O**nly **L**ook **A**t **C**oefficien**T**s) để giải quyết bài toán phân vùng thực thể bằng cách tận dụng *kiến trúc tích chập đầy đủ* (fully convolutional architecture). Bằng cách này, mô hình sẽ không còn phụ thuộc *mạng đề xuất khu vực* (region proposal networks) và *phương pháp anchor* (anchor-based methods), từ đó giúp cải thiện đáng kể thời gian dự đoán mà vẫn có thể bảo toàn độ chính xác. Đây là một phương pháp mạnh mẽ được sử dụng phổ biến để giải quyết các bài toán thị giác máy tính. Để triển khai ý tưởng này, chúng tôi chia bài toán phân vùng thực thể thành hai tác vụ song song: (1) sinh ra một tập các *mặt nạ mẫu* (prototype masks) và (2) dự đoán các *hệ số mặt nạ* (mask coefficients) cho mỗi thực thể. Sau đó, kết hợp tuyến tính mặt nạ mẫu với hệ số mặt nạ để tạo ra *mặt nạ thực thể* (instance masks) cho từng đối tượng trong ảnh. Cuối cùng, chúng tôi đề xuất Fast *NMS* (Non Maximum Suppression), một phương pháp thay thế NMS tiêu chuẩn, với mục đích làm giảm nhẹ hiệu suất để cải thiện thời gian thực thi. |
| **GIỚI THIỆU**  Phân vùng thực thể là một bài toán con quan trọng trong Thị giác Máy tính, được ứng dụng rộng rãi ở nhiều lĩnh vực như truy vết đối tượng, robot học, ảnh y khoa,… Bài toán phân vùng thực thể nhận đầu vào là một bức ảnh. Đầu ra là một tập các đối tượng đã được phân vùng, trong đó mỗi đối tượng được biểu diễn bằng một *khung giới hạn* (bounding boxes) chỉ ra vị trí của đối tượng trên ảnh và một *mặt nạ phân vùng* (segmentation masks) tương ứng gồm các giá trị nhị phân giúp xác định đối tượng ở cấp độ điểm ảnh.  Các cách tiếp cận SOTA (State-of-the-art) trong bài toán phân vùng thực thể như Mask R-CNN [1] hay FCIS [2] được xây dựng trực tiếp từ những phương pháp của bài toán nhận dạng đối tượng như Faster R-CNN [3] và R-FCN [4] , ưu tiên độ chính xác hơn là thời gian thực thi. Các *phương pháp phát hiện đối tượng 1 giai đoạn* (One-stage object detectors) như YOLO ([5], [6]) hay SSD [7] thực thi nhanh hơn các phương pháp phát hiện đối tượng 2 giai đoạn như Faster R-CNN [3] do lượt bỏ đi giai đoạn thứ 2 và bù đắp phần hiệu suất mất mát bằng nhiều cách khác nhau.  Trong đề tài này, chúng tôi đề xuất một phương pháp mới với tên gọi là YOLACT, một quy trình phân vùng thực thể thời gian thực với 2 tác vụ song song. Sau đó, kết hợp tuyến tính 2 đầu ra kể trên để cho ra mặt nạ kết quả cuối cùng. Cụ thể:  Đầu vào:  A bench in a park  Description automatically generated  Đầu ra:  A bench in a park  Description automatically generated |
| **MỤC TIÊU**   * Giảm thời gian thực thi: Tiến hành thử nghiệm song song 2 tác vụ sinh mặt nạ mẫu và dự đoán hệ số mặt nạ. Sau đó, thử nghiệm các cách kết hợp 2 đầu ra kể trên sao cho tối ưu hóa cho quá trình thực thi thời gian thực. * Cải thiện độ chính xác: Tiến hành nghiên cứu lại cách bảo toàn lượng thông tin mất mát ở các lớp pooling mà những phương pháp SOTA sử dụng. * Mang tính khái quát: Ý tưởng sinh mặt nạ mẫu và hệ số mặt nạ có thể được áp dụng vào đa số các phương pháp phân vùng thực thể. |
| **NỘI DUNG VÀ PHƯƠNG PHÁP**  Quy trình phân vùng thực thể của YOLACT sẽ bao gồm 3 tác vụ với (1) và (2) được thực hiện song song:   * (1) Sinh mẫu: có nhiệm vụ dự đoán một tập các mặt nạ mẫu khả dĩ trong phạm vi toàn bộ bức ảnh đầu vào. Thực hiện điều này thông qua việc nghiên cứu kiến trúc *mạng tích chập đầy đủ* (FCN – Fully Convolution Network) và cách sử dụng thành phần mạng sinh mẫu có sẵn ở các công trình trước ví dụ như FPN [8] . * (2) Hệ số mặt nạ: Thêm 1 tác vụ chạy song song ở bộ phận dự đoán như trong các phương pháp anchor, đó là dự đoán thêm các hệ số mặt nạ tương ứng với mỗi mẫu ở (1). * (3) Kết hợp tuyến tính ma trận mặt nạ mẫu (1) và ma trận hệ số mặt nạ (2) cho các đối tượng đã được duyệt qua hàm NMS. Có thể sử dụng phép nhân ma trận và thông qua một hàm phi tuyến để tạo ra mặt nạ thực thể cho từng đối tượng có trên ảnh.   Đề xuất thêm hàm Fast NMS nhằm lượt bớt các khung giới hạn bị trùng lắp. Điểm khác biệt so với NMS tiêu chuẩn là quá trình giữ lại và bỏ đi các đối tượng bị trùng lắp được thực hiện song song giúp tiết kiệm thời gian. Ý tưởng này được triển khai trên GPU và cần được nghiên cứu thêm.  Ý tưởng thiết kế mô hình dựa trên FPN [8], ResNet-101 [9] và RetinaNet [10].  Quá trình huấn luyện:   * Sử dụng 3 độ lỗi để huấn luyện mô hình: độ lỗi phân lớp, *độ lỗi hồi quy khung* (box regression loss) và độ lỗi mặt nạ. * Sử dụng tập dữ liệu ImageNet [11] . * Tận dụng khả năng tính toán song song của GPU. |
| **KẾT QUẢ MONG ĐỢI**   * Thời gian thực thi được rút ngắn: Các tác vụ được thực hiện song song cùng với bước kết hợp tuyến tính đơn giản. * Mặt nạ kết quả bao phủ đúng thực thể: Thông tin không gian bị mất mát ở các tầng pooling đã được bảo toàn. * Có tính khái quát: Ý tưởng triển khai được trên hầu hết những phương pháp thuộc phạm vi bài toán nhận dạng đối tượng. |
| **TÀI LIỆU THAM KHẢO**  [1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, “Mask R-CNN.” arXiv, Jan. 24, 2018. Accessed: Jul. 16, 2023. [Online]. Available: http://arxiv.org/abs/1703.06870  [2] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei, “Fully Convolutional Instance-aware Semantic Segmentation.” arXiv, Apr. 10, 2017. Accessed: Jul. 16, 2023. [Online]. Available: http://arxiv.org/abs/1611.07709  [3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” arXiv, Jan. 06, 2016. Accessed: Jul. 16, 2023. [Online]. Available: http://arxiv.org/abs/1506.01497  [4] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun, “R-FCN: Object Detection via Region-based Fully Convolutional Networks,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2016. Accessed: Jul. 16, 2023. [Online]. Available: https://papers.nips.cc/paper\_files/paper/2016/hash/577ef1154f3240ad5b9b413aa7346a1e-Abstract.html  [5] Joseph Redmon and Ali Farhadi, “YOLO9000: Better, Faster, Stronger.” arXiv, Dec. 25, 2016. Accessed: Jul. 16, 2023. [Online]. Available: http://arxiv.org/abs/1612.08242  [6] Joseph Redmon and Ali Farhadi, “YOLOv3: An Incremental Improvement.” arXiv, Apr. 08, 2018. Accessed: Jul. 16, 2023. [Online]. Available: http://arxiv.org/abs/1804.02767  [7] Wei Liu *et al.*, “SSD: Single Shot MultiBox Detector,” 2016, pp. 21–37. doi: 10.1007/978-3-319-46448-0\_2.  [8] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, “Feature Pyramid Networks for Object Detection,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 936–944. doi: 10.1109/CVPR.2017.106.  [9] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition.” arXiv, Dec. 10, 2015. Accessed: Jul. 17, 2023. [Online]. Available: http://arxiv.org/abs/1512.03385  [10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal Loss for Dense Object Detection,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, Oct. 2017, pp. 2999–3007. doi: 10.1109/ICCV.2017.324.  [11] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL: IEEE, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848. |