Tạp chí Khoa học Trường Đại học Cần Thơ
Phần A: Khoa học Tự nhiên, Công nghệ và Môi trường
website: ctujsvn.ctu.edu.vn

DOI:10.22144/ctu.jvn.2022.087

THỰC NGHIỆM ĐÁNH GIÁ YOLOX CHO BÀI TOÁN PHÁT HIỆN ĐỐI TƯỢNG TÀI LIỆU

Huỳnh Viết Tuấn Kiệt*, Nguyễn Văn Toàn, Nguyễn Trọng Thuận, Võ Duy Nguyên và Nguyễn Tấn Trần Minh Khang

Trường Đại học Công nghệ Thông Tin – Đại học Quốc gia Thành phố Hồ Chí Minh

**Người chịu trách nhiệm về bài viết: Huỳnh Viết Tuấn Kiệt (email: 20521494@gm.uit.edu.vn)*

Thông tin chung:
Ngày nhận bài: 05/01/2022
Ngày nhận bài sửa: 11/02/2022
Ngày duyệt đăng: 15/02/2022

Title:
Empirical evaluation of YOLOX for document object detection

Từ khóa:
Phát hiện đối tượng tài liệu, Phát hiện đối tượng trang, Phát hiện tài liệu tiếng Việt, Không có Anchor

Keywords:
Anchor-free, Document Object Detection, Page Object Detection, Vietnamese Document Detection

ABSTRACT

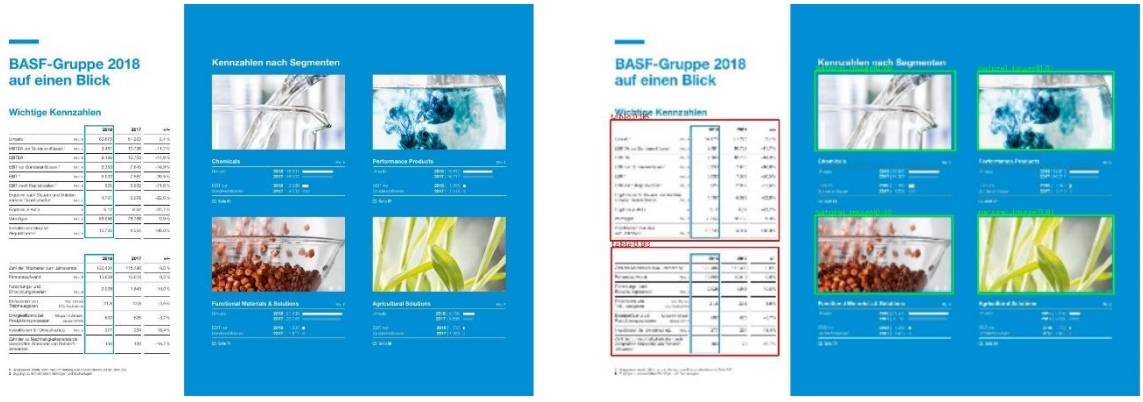
In the few recent decades, the rapidly increasing digitalization of image documents, accurate information extraction has been an important research area of the document analysis community. Many research works have been conducted on element-based approach for document classification. In this paper, the objective addresses the POD (Page Object Detection) problem – detecting objects that appear in document pages, by evaluating 2 datasets: IIIT-AR-13K and UIT-DODV as the benchmark for the YOLOX method. YOLOX achieved 69,0% mAP on the UIT-DODV dataset and 66,9% mAP on the IIIT-AR-13K dataset. Compared to the highest result of the previous state-of-the-art of one-stage detector - YOLOv4x-mish, on the UIT-DODV dataset, YOLOX surpassed by 2,90% mAP. YOLOX is significantly lower in IIIT-AR-13K than in previously announced two-stage approaches. Furthermore, this research provided an analysis on the effectiveness of the state-of-the-art method YOLOX on the POD problem, which will become a premise for future researches.

TÓM TẮT

Trong vài thập kỷ qua, với sự gia tăng nhanh chóng trong việc số hóa các hình ảnh tài liệu, việc trích xuất thông tin chính xác là một trong những hướng nghiên cứu quan trọng. Với sự phát triển của phát hiện đối tượng, nhiều nghiên cứu ra đời hướng đến việc phân loại tài liệu dựa trên nhiều thành phần của trang tài liệu đó. Mục tiêu của nghiên cứu này là đề cập đến bài toán POD (Page Object Detection) – phát hiện đối tượng xuất hiện trong trang tài liệu thông qua đánh giá 2 bộ dữ liệu IIIT-AR-13K và UIT-DODV dựa theo phương pháp YOLOX. YOLOX đạt kết quả 69,0% mAP, tốt hơn 2,90% so với kết quả mô hình one-stage cao nhất – YOLOv4-mish được công bố trên bộ dữ liệu UIT-DODV. Trong khi ở IIIT-AR-13K, YOLOX đạt được 66,9% mAP và thấp hơn nhiều so với các phương pháp two-stage đã công bố trước đó. Bên cạnh, những phân tích về độ hiệu quả của phương pháp state-of-the-art YOLOX cho bài toán POD cũng được cung cấp, là tiền đề cho những nghiên cứu tiếp theo trong tương lai.

1. GIỚI THIỆU

Qua nhiều thế kỷ, tài liệu giấy vẫn là công cụ chính để tạo nên sự tiến bộ lâu dài của loài người. Ngày nay, hầu hết thông tin vẫn được ghi lại, lưu trữ và phân phối dưới dạng giấy. Việc sử dụng máy tính để chỉnh sửa tài liệu và sự ra đời của bộ xử lý văn bản vào cuối những năm 1980 đã thay thế tài liệu truyền thống (giấy, tờ báo, sách,...), tài liệu số (WORD, PDF...) xuất hiện dẫn đến sự gia tăng nhanh chóng trong việc số hóa tài liệu và cải thiện đáng kể khả năng tiếp cận dữ liệu và được lưu trữ thông qua các dịch vụ điện toán đám mây để thuận tiện cho truy cập, tìm kiếm, sao lưu tài liệu (Marinai, 2008). Bên cạnh những thuận lợi đó là khối tài liệu khổng lồ dẫn đến việc truy cập trở nên khó khăn hơn. Với các thuật toán phát hiện đối tượng dựa trên học sâu gần đây trong lĩnh vực thị giác máy tính, một lượng phương pháp đáng kể được phát triển đã xây dựng mô hình phát hiện các đối tượng trang đồ họa trong hình ảnh tài liệu như một vấn đề phát hiện đối tượng (Bhatt et al., 2021). Dựa theo sự phát triển đó, phạm vi của nghiên cứu này là bài toán phát hiện các thành phần quan trọng xuất hiện trong trang tài liệu như "Caption", "Table", "Figure", "Formula", ... "Document Image Understanding" (Gao et al., 2017) là một nghiên cứu quan trọng được thực hiện với nhiều vấn đề thách thức, đang nhận được sự quan tâm ngày càng nhiều không chỉ từ các cộng đồng phân tích và ghi nhận tài liệu. Bài toán phát hiện đối tượng trang trong hình ảnh tài liệu (Nguyen et al., 2018; Long và ctv., 2020; Le et al., 2021; Nguyen et al., 2022) vẫn là một thách thức vì các đối tượng trang rất đa dạng về quy mô và tỷ lệ khung hình, và một đối tượng có thể chứa các thành phần gần như tách rời nhau. Do đó, việc rút trích thông tin từ hình ảnh của tài liệu là vô cùng cần thiết, nhiều phương pháp máy học ra đời trong tương lai sẽ giúp con người dễ dàng tìm kiếm những tài liệu cần thiết và tránh mất nhiều thời gian. Hướng giải quyết của những phương pháp được đề xuất trước đây chưa mang lại độ chính xác cao và tốn nhiều thời gian để xử lí. Dựa theo nhiều công trình nghiên cứu khoa học trước đó và đứng trên góc nhìn của bài toán phát hiện đối tượng, bài báo này là bài toán phát hiện đối tượng trang (Hình 1) được đề cập thông qua việc đánh giá 2 bộ dữ liệu IIIT-AR-13K (Mondal et al., 2020) và UIT-DODV (Dieu et al., 2021) dựa theo phương pháp YOLOX (Ge et al., 2021). Sơ lược về các bộ dữ liệu sử dụng trong bài toán, bộ dữ liệu IIIT-AR-13K (Mondal et al., 2020) chứa tổng cộng hơn 13,000 hình ảnh trang được chú thích với các đối tượng thuộc 5 danh mục phổ biến khác nhau gồm "Table", "Figure", "Natural image", "Logo" và "Signature". Mondal et al. (2020) đã tạo bộ dữ liệu này theo cách thủ công bằng việc sưu tầm rộng rãi từ các bài báo cáo, tạp chí khoa học, bài nghiên cứu khắp nơi trên thế giới và đây là một trong những bộ dữ liệu lớn nhất trong lĩnh vực phát hiện đối tượng trang đồ họa (Bhatt et al., 2021). Trong khi đó, bộ dữ liệu UIT-DODV (Dieu et al., 2021) là bộ dữ liệu về tài liệu tiếng Việt đầu tiên với các đối tượng của hình ảnh đầu vào bao gồm "Caption", "Table", "Figure" và "Formula". Đặc điểm của UIT-DODV là các trang tài liệu tiếng Việt, do đó mang lại nhiều tính mới mẻ. Ví dụ như cách trình bày các đối tượng ngữ nghĩa tạo ra nhiều khó khăn trong việc rút trích đặc trưng các thông tin, các công thức không chỉ là công thức toán học bình thường mà còn ở các dạng không thuộc toán học.



a. Đầu vào của bài toán **b. Đầu ra của bài toán**

Hình 1. Minh họa bài toán phát hiện đối tượng trang tài liệu. Đầu vào (a) là trang hình ảnh tài liệu chứa các đối tượng hoặc có thể không chứa đối tượng nào và đầu ra (b) là các vị trí của đối tượng có thể có trong ảnh được bao bọc bằng các bounding box

Thực nghiệm và đánh giá phương pháp YOLOX trên hai bộ dữ liệu IIIT-AR-13K cùng với UIT-DODV đã thu được những kết quả khá tốt. Đánh giá trên bộ dữ liệu UIT-DODV thì mô hình đạt được 69,0% mAP, kết quả này cao hơn 2,90% so với YOLOv4x-mish là mô hình one-stage đạt kết quả cao nhất được Dieu et al. (2021) công bố trước đó. Đánh giá trên bộ dữ liệu IIIT-AR-13K, mô hình đạt được 66,9% mAP, kết quả này thấp hơn so với các phương pháp two-stage Faster-RCNN và Mask-RCNN đã được công bố (Mondal et al., 2020).

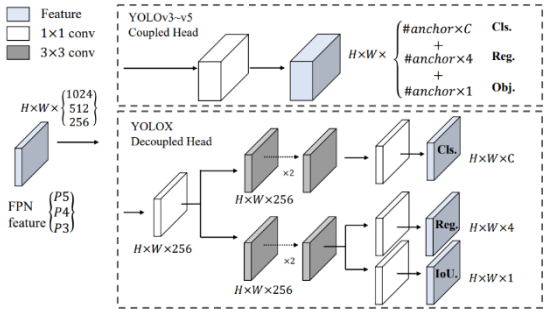
2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Các nghiên cứu liên quan

Trọng tâm của nghiên cứu hướng đến đánh giá các phương pháp phát hiện đối tượng theo thời gian thực, vì vậy các phương pháp phát hiện đối tượng one-stage sẽ được quan tâm và đề cập đến trong bài viết. RetinaNet (Lin et al., 2017) là một mô hình giải quyết được vấn đề mất cân bằng trong phân phối giữa "foreground" và "background" trong các bài toán phát hiện một giai đoạn bằng cách sử dụng hàm "Focal Loss" thay cho "Cross Entropy". EfficientDet (Tan et al., 2020) sử dụng một số đặc trưng tối ưu hóa và chỉnh sửa backbone, chẳng hạn như sử dụng BiFPN và phương pháp chia tỷ lệ kết hợp giúp chia tỷ lệ đồng nhất độ phân giải, độ sâu và chiều rộng cho tất cả các backbone, "Feature Networks" và dự đoán box/class cùng lúc. YOLOv4 (Bochkovskiy et al., 2020) nhanh hơn gấp đôi so với EfficientDet (Tan et al., 2020) nhưng đạt hiệu suất tương đương, YOLOv4 bao gồm CSPDarknet-53 làm backbone, kim tự tháp không gian tích hợp module bổ sung, mạng tổng hợp đường dẫn (PANet) và đầu YOLOv3. YOLOv4x-mish là phiên bản bổ sung với một số thay đổi so với YOLOv4. YOLOv4x-mish thay đổi giai đoạn CSPDarknet đầu tiên của phần backbone thành Darknet-53 ban đầu giúp cân bằng tốc độ và độ chính xác. YOLOF (Chen et al., 2021) là một cải tiến từ các phương pháp thuộc họ YOLO, đồng thời phương pháp này cũng cải thiện được những nhược điểm từ RetinaNet. YOLOF xem xét lại đặc trưng mạng kim tự tháp (Pyramids Networks) cho các máy dò một giai đoạn (one-stage detectors) và chỉ ra rằng sự thành công của "Feature Pyramid Networks" là do giải pháp chia để trị (devide-and-conquer) của nó cho vấn đề tối ưu hóa trong phát hiện đối tượng chứ không phải là sự kết hợp tính năng đa quy mô (multi-scale feature fusion). Bên cạnh, trong những năm gần đây, "Anchor-free" là một trong những hướng nghiên cứu mới nhận được sự quan tâm. Đặc biệt, phương pháp YOLOF đã được Nguyen et al. (2021) áp dụng cho bài toán phát hiện đối tượng trang. CenterNet (Duan et al., 2019) là một trong những phương pháp one-stage áp dụng kĩ thuật "Anchor-free", trong phương pháp này, bài toán phát hiện đối tượng (Object Detection) được đưa về bài toán tìm điểm đặc trưng (Keypoint Estimation), từ đó cũng suy ra kích thước và tính toán được bounding box cho bài toán phát hiện vật.

2.2. Phương pháp

YOLOX là phương pháp được đề xuất bởi Ge et al. (2021) và có sự hậu thuẫn của công ty công nghệ Megvii. YOLOX là phiên bản không có mỏ neo (Anchor-free) của YOLO, với thiết kế đơn giản hơn nhưng hiệu suất tốt hơn. Cụ thể YOLOX là một phiên bản cải tiến dựa trên YOLOv3 baseline (Redmon & Farhadi, 2018). Các thay đổi mô hình lớn nhất bao gồm việc loại bỏ các "Anchor box" và áp dụng các kỹ thuật nâng cao hiện tại để phát hiện đối tượng. Trong quá trình phát hiện đối tượng dựa trên "Anchor box", Ge et al. (2021) cho thấy việc đặt quá nhiều "Anchor box" giữa hình ảnh sẽ để lại nhiều nhược điểm. Chẳng hạn như muốn tối ưu hiệu suất phải tiến hành phân tích phân cụm để xác định ra được "Anchor box" tối ưu, các cấu hình thu được thường theo miền cụ thể và không thể tổng quát hóa cho các bộ dữ liệu khác. Bên cạnh, "Anchor box" làm tăng độ phức tạp của các đầu và số lượng các dự đoán cho mỗi hình ảnh. Cơ chế "Anchor-free" làm giảm đáng kể số lượng các thông số thiết kế cần điều chỉnh heuristic và nhiều thủ thuật liên quan (ví dụ: Anchor Clustering, Grid Sensitive) để có hiệu suất tốt, đặc biệt là giai đoạn huấn luyện và giải mã của nó đơn giản hơn đáng kể (Ge et al., 2021).



Hình 2. Minh họa sự khác nhau giữa coupled head trong kiến trúc YOLOv3 và decoupled head trong kiến trúc YOLOX được Ge et al. (2021) đề xuất (thêm hai nhánh song song với hai lớp chuyển đổi 3 × 3, mỗi nhánh cho các nhiệm vụ phân loại và hồi quy tương ứng) ***(Ge et al., 2021)***

Các phiên bản YOLO từ trước (Redmon et al., 2016; Redmon & Farhadi, 2018; Bochkovskiy et al., 2020) vẫn sử dụng đầu ghép (coupled head) trong trích xuất tính năng. Hai thí nghiệm phân tích của Ge et al. (2021) trong nghiên cứu YOLOX chỉ ra rằng đầu phát hiện được ghép nối (coupled head) có thể gây hại cho hiệu suất. Nhằm giải quyết những vấn đề trên, YOLOX đã tách đầu dò YOLO thành các kênh có tính năng riêng biệt để thực hiện nhiệm vụ phân loại và hồi quy. Thay thế đầu của YOLO bằng một đầu tách rời (decoupled head) để cải thiện thời gian hội tụ huấn luyện và độ chính xác của mô hình (Hình 2) (Ge et al., 2021). Cuối cùng để giảm thiểu sự mất cân bằng cực độ giữa việc lấy mẫu tích cực/ tiêu cực khi huấn luyện, thay vì chỉ chọn 1 mẫu tích cực tại vị trí trung tâm cho mỗi đối tượng, giải pháp chỉ định trung tâm 3×3 là các mẫu tích cực. Chiến lược này được gọi là "center sampling" trong FCOS (Tian et al., 2019).

Hình ảnh đầu vào của cả 2 bộ dữ liệu thực nghiệm có sự khác nhau về kích thước. Do đó, trước khi huấn luyện mô hình, các hình ảnh được điều chỉnh về cùng một kích thước. Kích thước hình ảnh điều chỉnh và các tham số khác của mô hình được áp dụng theo mặc định của YOLOX được MMDetection (Chen et el., 2019) cung cấp.

2.3. Bộ dữ liệu

Nghiên cứu được thực hiện trên 2 bộ dữ liệu IIIT-AR-13K (Mondal et al., 2020) và UIT-DODV (Dieu et al., 2021). Những bộ dữ liệu này chứa nhiều loại nhãn chính và có sự đa dạng, phức tạp về ngôn ngữ, đặc biệt là tiếng Việt, tạo nên nhiều thách thức cho mô hình khi huấn luyện. IIIT-AR-13K là bộ dữ liệu được thu thập từ các trang báo với nhiều ngôn ngữ khác nhau (Anh, Pháp, Nhật,...) trong hơn 10 năm qua và đã được công bố vào năm 2020. Bộ dữ liệu này chứa khoảng 13.000 hình ảnh tài liệu được chia làm 3 tập dữ liệu bao gồm: 9.333 hình ảnh cho tập Train, 1.955 hình ảnh cho tập Validation, và 2.120 hình ảnh cho tập Test, tất cả được thống kê trong hình Bảng 1.

Bảng 1. Thống kê bộ dữ liệu IIIT-AR-13K

Object	Train	Validation	Test	Total
Table	11.163	2.222	2.596	15.981
Figure	2.004	481	463	2.948
Natural image	1.987	438	455	2.880
Logo	379	67	135	581
Signature	420	108	92	620

(Mondal et al., 2020)

Nghiên cứu này hướng đến đánh giá bài toán dựa trên sự đa đạng và khác nhau về nhiều ngôn ngữ, do đó, UIT-DODV là bộ dữ liệu tiếp theo được lựa chọn thực nghiệm. Bộ dữ liệu UIT-DODV khá mới lạ khi đây là bộ dữ liệu hoàn toàn bằng tiếng Việt đầu tiên với các hình ảnh đầu vào là sự kết hợp của các đối tượng "Caption", "Table", "Figure" và "Formula". Bộ dữ liệu được Dieu et al. (2021) công bố vào năm 2021 chứa tổng cộng 2.394 hình ảnh được chia thành 3 tập dữ liệu bao gồm: 1.440 ảnh cho tập Train, 234 ảnh cho tập Validation, 720 ảnh cho tập Test và được thống kê trong Bảng 2.

Chọn một bộ dữ liệu lớn và một bộ dữ liệu khá mới lạ trong nghiên cứu sẽ làm tăng độ phức tạp và tạo ra những khó khăn cho mô hình khi phát hiện đối tượng, đổi lại nghiên cứu này sẽ đưa ra được những đánh giá tốt hơn về phương pháp YOLOX.

Bảng 2. Thống kê bộ dữ liệu UIT-DODV

Object	Train	Validation	Test	Total
Table	1.929	149	548	2.626
Figure	1.143	212	678	2.033
Caption	2.106	334	1.174	3.614
Formula	1.349	83	330	1.762

(Dieu et al., 2021)

2.4. Cấu hình thực nghiệm

Thiết lập thực nghiệm trong nghiên cứu này trên cấu hình 2x GPU RTX 2080Ti và sử dụng cấu hình mặc định của YOLOX được framework MMDetection (Chen et al., 2019) cung cấp.

2.5. Chỉ số đánh giá

Sau giai đoạn thực nghiệm là quá trình đánh giá mô hình YOLOX trên 2 bộ dữ liệu IIIT-AR-13K và UIT-DODV, độ đo mAP (Mean Average Precision) (Lin et al., 2014) được sử dụng để đánh giá trong nghiên cứu. Kết quả có được trên các độ đo AP_{50} và AP_{75} tương ứng với ngưỡng IoU (Intersection over Union) lần lượt là 0,5 và 0,75. Lần lượt tính AP (average precision) của mỗi lớp xuất hiện trong trang với các ngưỡng IoU khác nhau, từ đó tính trung bình để lấy AP của lớp đó. Sau khi có AP của tất cả các lớp, kết quả cuối cùng cho mô hình sẽ được đưa ra bằng việc tính trung bình để có được độ đo mAP.

3. KẾT QUẢ VÀ THẢO LUẬN

Báo cáo kết quả thực nghiệm YOLOX (Ge et al., 2021) trên 2 bộ dữ liệu IIIT-AR-13K (Mondal et al., 2020) và UIT-DODV (Dieu et al., 2021) được trình bày trong Bảng 3.

Trên bộ dữ liệu IIIT-AR-13K, mô hình đề xuất về cơ bản không gặp nhiều khó khăn trong việc phát hiện các đối tượng "Table" và "Natural image", mô

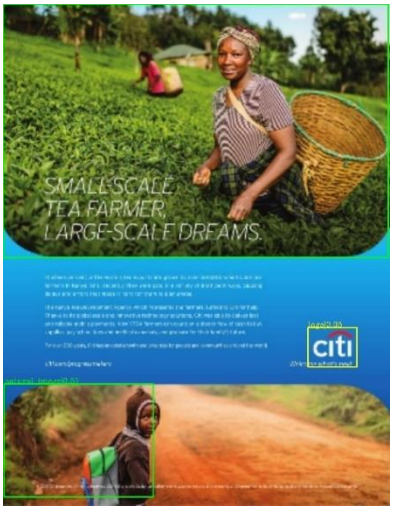
hình phân loại khá chính xác đối với hai đối tượng trên với kết quả 90,3% AP cho lớp "Table" và 86,0% AP cho lớp "Natural image", mặc dù các chi tiết thuộc lớp "Natural image" đa dạng về khung hình và màu sắc nhưng kết quả đánh giá đã chỉ ra mô hình phát hiện rất chính xác và được minh chứng đại diện một trang ở Hình 3. Một số trang tài liệu dự đoán chưa hoàn toàn chính xác về hai đối tượng trên nhưng phần lớn không mắc lỗi phát hiện sai mà do lỗi các bounding box chưa bao phủ hoàn toàn đối tượng (Hình 4a). Lớp "Figure" đạt 66,4% AP và lớp "Signature" đạt 58,1% AP, hai lớp này đạt hiệu suất phát hiện ở mức tương đối và vẫn có những sai sót đáng quan tâm như việc bỏ sót đối tượng và nhầm lẫn với các đối tượng khác, với chỉ hơn 600 đối tượng "Signature" có trong cả bộ dữ liệu thì kết quả đạt được cũng có thể chấp nhận đối với lớp này. Theo Bảng 1, có thể thấy số lượng đối tượng "Logo" và "Signature" xuất hiện trong bộ dữ liệu IIIT-AR-13K khá ít và gần như bằng nhau nhưng kết quả phát hiện của lớp "Logo" chỉ đạt 33,9% AP so với 58,1% AP của "Signature" đã đề cập ở trên. Bên cạnh việc số lượng đối tượng không đủ lớn, có thể lí giải "Logo" là đối tượng được biểu diễn bằng rất nhiều hình dạng, hình ảnh khác nhau cho nên dễ gây nhầm lẫn với các đối tượng "Signature" và "Figure" (Hình 4b). Ngoài những vấn đề phát sinh trong việc phát hiện đối tượng như vấn đề bỏ sót đối tượng cũng chưa được cải thiện tối ưu cho bài toán này, rất nhiều đối tượng khá rõ ràng nhưng bị bỏ qua đặc biệt các đối tượng "Logo" và "Natural image" (Hình 4d). Trên IIIT-AR-13K, chênh lệch khá lớn khi tăng ngưỡng IoU từ 0,5 lên 0,75 trong thực nghiệm và đánh giá, kết quả đo được trên bộ dữ liệu IIIT-AR-13K giảm từ 85,3% độ đo AP_{50} xuống còn 74,1% AP_{75}. Với nhiều sai sót cũng như hiệu suất đạt được trên lớp "Logo" còn khá thấp đã ảnh hưởng đến kết quả chung của mô hình, nhưng nhìn chung mô hình vẫn đạt được hiệu suất khá tốt với 66,9% mAP sau khi kết hợp tất cả các lớp lại. So với các phương pháp two-stage đã được đánh giá và công bố (Mondal et al., 2020), đối tượng "Figure" và "Signature" ở bộ dữ liệu IIIT-AR-13K cho kết quả tốt hơn nhiều trên hai phương pháp Faster-RCNN và Mask-RCNN trong khi đối tượng "Logo" vẫn còn tương đối thấp, do đó việc tăng cường đối tượng "Logo" có thể là một giải pháp được lựa chọn để cải thiện kết quả cuối cùng của mô hình trên mọi phương pháp thực nghiệm. Tóm lại, trên bộ dữ liệu IIIT-AR-13K, phương pháp one-stage YOLOX được đề xuất thực nghiệm trong nghiên cứu này cho hiệu suất phát hiện đối tượng thấp hơn nhiều khi so sánh với các phương pháp two-stage đã được công bố (Mondal et al., 2020) gồm có Faster-RCNN đạt 78,96% mAP và Mask-RCNN đạt 82,2% mAP.

Hình 3. Trang tài liệu trích trong tập test của bộ dữ liệu IIIT-AR-13K. Các bounding box màu xanh lục biểu thị cho lớp Natural image, xanh dương là Figure, màu vàng là Logo

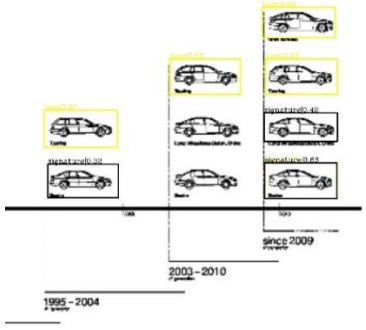
Đối với bộ dữ liệu UIT-DODV, mô hình cũng có được những điểm mạnh và những điểm hạn chế khá tương đồng như bộ dữ liệu IIIT-AR-13K. Kết quả ở Bảng 3 cho thấy một lần nữa mô hình YOLOX không gặp khó khăn trong việc phát hiện những đối tượng "Table" với 90,5% AP vượt trội so với các đối tượng khác trong bộ dữ liệu. Lớp "Figure" cũng cho kết quả ổn định với 82,0% AP, nhìn chung các đối tượng "Figure" có những đặc trưng riêng biệt so với các đối tượng còn lại. Tuy nhiên, không ít những trang tài liệu đã bỏ sót đối tượng này và quá trình phát hiện có xảy ra nhầm lẫn với đối tượng "Table" (Hình 5a), cho nên đây cũng là một phần lí do giải thích cho kết quả chưa được cao nhất ở lớp "Figure". Đạt AP ở mức khá với 63,5% AP là lớp "Caption", các đối tượng "Caption" thường được phân loại khá chính xác, nhưng xuất hiện nhiều trang tài liệu lại bỏ sót các đối tượng này và Hình 5a là một trong những trang điển hình. Vấn đề lớn nhất của bài toán trên tập dữ liệu UIT-DODV là lớp "Formula" khi chỉ đạt 40,1% AP, mô hình chưa phân biệt rõ ràng công thức tính toán và những câu văn có chứa kí tự số dẫn đến lỗi bỏ sót đối tượng, bounding box bao chưa chính xác hay phát hiện nhầm đối tượng, điển hình là trang tài liệu được đề cập ở Hình 5d khi đầu vào không chứa bất kì công thức tính toán nào trong khi đầu ra lại dự đoán được rất nhiều công thức, đây thực sự là vấn đề nghiêm trọng của mô hình. Thống kê trong Bảng 3 cho thấy kết quả chênh lệch khi tăng ngưỡng IoU từ 0,5 lên 0,75 rất lớn, từ 87,1% độ đo

AP_{50} xuống còn 57,7% AP_{75}. Với những sai sót đã được thảo luận và chỉ ra ở Hình 5, có thể thấy kết quả cuối cùng của mô hình sẽ không đạt được quá cao nhưng cũng tương đối ổn định với 69,0% mAP sau khi đã tính trung bình của tất cả các lớp, so với mô hình baseline của bộ dữ liệu UIT-DODV thì vẫn thấp hơn. Tuy nhiên, kết quả đánh giá của mô hình có thể vẫn sẽ được cải tiến tốt hơn trong những nghiên cứu khác liên quan trong tương lai. Ngoài ra, khi đánh giá trên bộ dữ liệu UIT-DODV, YOLOX cũng đã cho thấy kết quả vượt trội so với các phương pháp thuộc họ YOLO trước đó đã được Dieu et al., (2021) công bố. Cụ thể, YOLOX đã lần lượt cao hơn 3,2% và 2,9% so với 2 phương pháp trước đó là YOLOv4 và YOLOv4x-mish. Các kết quả này cho thấy sự hiệu quả của phương pháp one-stage YOLOX trong việc giải quyết bài toán phát hiện đối tượng trên tài liệu tiếng Việt dạng ảnh.

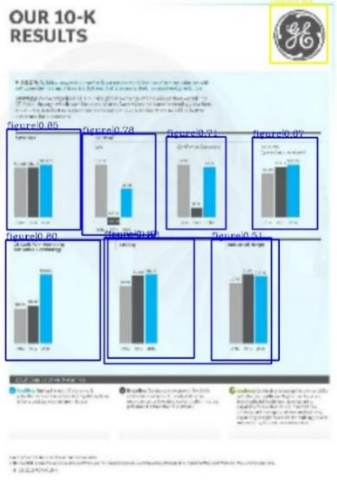
Đánh giá tổng quan dựa trên những phân tích trên ở cả hai bộ dữ liệu, mô hình YOLOX đạt kết quả phát hiện trên trang tài liệu dạng ảnh tương đối ổn định, cơ bản nhất vẫn đạt được hiệu suất cao với các đối tượng "Table", vẫn xảy ra vấn đề bỏ sót đối tượng xuất hiện rõ ràng trong trang (Hình 4d), (Hình 5a) và xảy ra tình trạng overlapping các bounding box (Hình 4c, Hình 5b, Hình 5c).



a. Bao không hết đối tượng



b. Dự đoán sai đối tượng



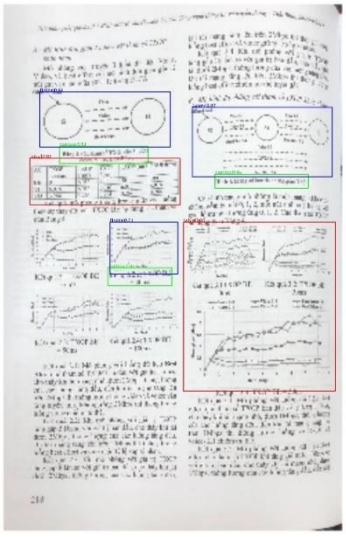
c. Overlapping bounding box

d. Bỏ sót đối tượng

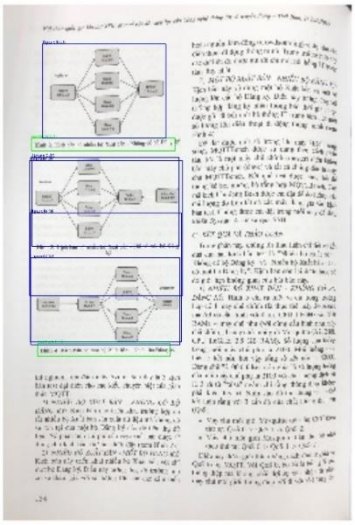
Hình 4. Trực quan những trang tài liệu dự đoán chưa tốt sau khi thực nghiệm trên bộ dữ liệu IIIT-AR-13K. Các bounding box màu đỏ biểu thị cho đối tượng Table, màu xanh lục là Natural image, xanh dương là Figure, màu vàng là Logo, màu đen là Signature

Bảng 3. So sánh kết quả thực nghiệm của các phương pháp phát hiện đối tượng một giai đoạn trên hai bộ dữ liệu IIIT-AR-13K và UIT-DODV

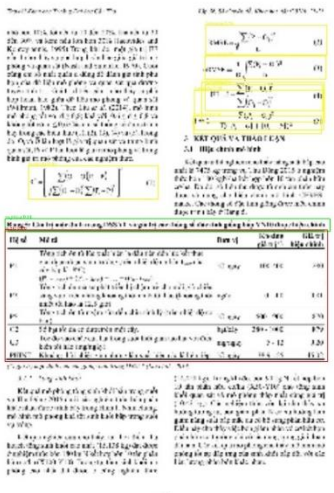
Method	Dataset	Table	Figure	Caption	Formula	Natural image	Logo	Signature	AP_{50}	AP_{75}	mAP
YOLOv4 (Dieu et al.)	UIT-DODV	84,2	78,0	60,8	40,2	-	-	-	90,2	75,2	65,8
YOLOv4x-mish (Dieu et al.)	UIT-DODV	82,0	75,7	61,3	**45,2**	-	-	-	**90,7**	**77,7**	66,1
YOLOF (Nguyen et al.)	IIIT-AR-13K	88,3	63,7	-	-	73,0	18,3	50,6	81,2	64,9	58,8
	UIT-DODV	86,4	74,4	31,4	31,8	-	-	-	79,3	59,5	56,0
YOLOX	IIIT-AR-13K	90,3	66,4	-	-	**86,0**	**33,9**	**58,1**	85,3	74,1	66,9
	UIT-DODV	**90,5**	**82,0**	**63,5**	40,1	-	-	-	87,1	75,7	**69,0**



a. Dự đoán sai và bỏ sót đối tượng



b. Overlapping bounding box



c. Overlapping bounding box

d. Đầu ra trang tài liệu

Hình 5. Trực quan những trang tài liệu dự đoán chưa tốt sau khi thực nghiệm trên bộ dữ liệu UIT-DODV. Các bounding box màu đỏ biểu thị cho đối tượng Table, màu xanh lục là Caption, xanh dương là Figure, màu vàng là Formula

4. KẾT LUẬN

Nghiên cứu lấy cảm hứng dựa trên thời đại của sự gia tăng nhanh chóng trong việc số các hình ảnh tài liệu, song song với đó là tính cần thiết của bài toán Page Object Detection. Nghiên cứu này đã giới thiệu phương pháp YOLOX và đánh giá trên hai bộ dữ liệu tài liệu bao gồm bộ dữ liệu IIIT-AR-13K và UIT-DODV. Cả hai bộ dữ liệu đều có những ưu điểm và có những thách thức riêng, kết quả đạt được cũng mang tính ổn định trên cả hai bộ dữ liệu. Các vấn đề đã được trình bày trong nghiên cứu góp phần mang lại nguồn cảm hứng và những đóng góp hữu ích cho các nghiên cứu liên quan. Trong tương lai, nhiều nghiên cứu khác sẽ được thực hiện để tiếp tục cải thiện hiệu suất của mô hình trên các bộ dữ liệu về tài liệu dạng ảnh.

LỜI CẢM TẠ

Nghiên cứu được thực hiện tại Phòng thí nghiệm Truyền thông Đa phương tiện (MMLab), Trường đại học Công nghệ Thông tin – Đại học Quốc gia Thành phố Hồ Chí Minh,

TÀI LIỆU THAM KHẢO

Bhatt, J., Hashmi, K. A., Afzal, M. Z., & Stricker, D. (2021). A Survey of Graphical Page Object Detection with Deep Neural Networks. *Applied Sciences, 11*(12), 5344. https://doi.org/10.3390/app11125344

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934.*

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., ... & Lin, D. (2019). MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155.*

Chen, Q., Wang, Y., Yang, T., Zhang, X., Cheng, J., & Sun, J. (2021). You only look one-level feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13039-13048).

Dieu, L. T., Nguyen, T. T., Vo, N. D., Nguyen, T. V., & Nguyen, K. (2021, September). Parsing Digitized Vietnamese Paper Documents. In *International Conference on Computer Analysis of Images and Patterns* (pp. 382-392). Springer, Cham. https://www.doi.org/10.1007/978-3-030-89128-2_37

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., & Tian, Q. (2019). Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6569-6578). https://doi.org/10.1109/ICCV.2019.00667

Gao, L., Yi, X., Jiang, Z., Hao, L., & Tang, Z. (2017, November). ICDAR2017 competition on page object detection. In 2017 *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 1417-1422). IEEE. https://www.doi.org/10.1109/ICDAR.2017.231

Ge, Z., Liu, S., Wang, F., Li, Z., & Sun, J. (2021). Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430.*

Le, H., Nguyen, T., Le, V., Nguyen, T. T., Vo, N. D., & Nguyen, K. (2021, December). Guided Anchoring Cascade R-CNN: An intensive improvement of R-CNN in Vietnamese Document Detection (2021). In *Proceedings of NAFOSTED Conference on Information and Computer Science (NICS)* (pp. 205-210). https://doi.org/10.1109/NICS54270.2021.9701510

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988). https://doi.org/10.1109/ICCV.2017.324Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740-755). Springer, Cham. https://doi.org/10.1007/978-3-319-10602-1_48

Long, D. P., Hiếu, N. T., Vi, N. T. T., Nguyên, V. D., & Khang, N. T. T. M. (2020). Phát hiện bảng trong tài liệu dạng ảnh sử dụng phương pháp định vị góc CornerNet. In *Proceedings of Fundamental and Applied Information Technology Research (FAIR).*

Marinai, S. (2008). Introduction to document analysis and recognition. In *Machine learning in document analysis and recognition* (pp. 1-20). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-76280-5_1

Mondal, A., Lipps, P., & Jawahar, C. V. (2020, July). IIIT-AR-13K: a new dataset for graphical object detection in documents. In *International Workshop on Document Analysis Systems* (pp. 216-230). Springer, Cham. https://doi.org/10.1007/978-3-030-57058-3_16

Nguyen, T. T., Nguyen, T. Q., Duong, L., Vo, N. D., & Nguyen, K. (2022). CDeRSNet: Towards High Performance Object Detection in Vietnamese Documents Images. In *International Conference on Multimedia Modelling (MMM).* https://doi.org/10.1007/978-3-030-98355-0_36

Nguyen, P., Ngo, L., Truong, T., Nguyen, T. T., Vo, N. D., & Nguyen, K. (2021, December). Page Object Detection with YOLOF. In *Proceedings of NAFOSTED Conference on Information and Computer Science (NICS)* (pp. 205-210). https://doi.org/10.1109/NICS54270.2021.9701449

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Tan, M., Pang, R., & Le, Q. V. (2020). Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10781-10790).

Tian, Z., Shen, C., Chen, H., & He, T. (2019). Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9627-9636).

Nguyen, D., Vo, Khanh-Duy Nguyen, Tam, V., Nguyen., & Nguyen, K. (2018, January). Ensemble of deep object detectors for page object detection. In *Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication* (pp. 1-6). https://doi.org/10.1145/3164541.3164644