

THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo: <https://youtu.be/C8iep3TsBYY>
- Link slides: https://github.com/Yangchann/CS519.O21.KHTN/blob/main/CS519_Slide.pdf

<ul style="list-style-type: none">• Họ và Tên: Trần Như Cẩm Nguyên• MSSV: 22520004 	<ul style="list-style-type: none">• Lớp: CS519.O21.KHTN• Tự đánh giá (điểm tổng kết môn): 8.5/10• Số buổi vắng: 0• Số câu hỏi QT cá nhân: 9• Link Github: https://github.com/cnmeow/CS519.O21_KHTN/• Mô tả công việc và đóng góp cá nhân:<ul style="list-style-type: none">○ Lên ý tưởng đề tài○ Viết phần tóm tắt, mục tiêu nghiên cứu và kết quả mong đợi○ Làm poster và chỉnh sửa video
<ul style="list-style-type: none">• Họ và Tên: Trần Thị Cẩm Giang• MSSV: 22520361 	<ul style="list-style-type: none">• Lớp: CS519.O21.KHTN• Tự đánh giá (điểm tổng kết môn): 8.5/10• Số buổi vắng: 1• Số câu hỏi QT cá nhân: 9• Link Github: https://github.com/Yangchann/CS519.O21_KHTN/• Mô tả công việc và đóng góp cá nhân:<ul style="list-style-type: none">○ Lên ý tưởng đề tài○ Viết phần giới thiệu, nội dung và phương pháp○ Làm slide và quay video YouTube

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI

HỢP NHẤT PHÁT HIỆN VÀ PHÂN TÍCH BỐ CỤC VĂN BẢN TRÊN BỘ DỮ LIỆU BẢNG HIỆU TIẾNG VIỆT PHÂN CẤP

TÊN ĐỀ TÀI TIẾNG ANH

UNIFIED TEXT DETECTION AND LAYOUT ANALYSIS ON HIERARCHICAL VIETNAMESE SIGNBOARD DATASET

TÓM TẮT

Trước đây, phát hiện văn bản trong ảnh ngoại cảnh (*scene text detection*) và phân tích bố cục tài liệu (*document layout analysis*) thường được coi là hai nhiệm vụ riêng biệt. Tuy nhiên, hai nhiệm vụ này có mối liên hệ chặt chẽ với nhau. Nghiên cứu này nhằm đánh giá hiệu suất của một số phương pháp SOTA hiện có và đề xuất phương pháp mới để hợp nhất hai nhiệm vụ đó.

Trong đề tài này, chúng tôi giới thiệu *Hierarchical Vietnamese Signboard Dataset*, một bộ dữ liệu được tạo ra nhằm phục vụ nghiên cứu về phát hiện văn bản và phân tích bố cục trên các bảng hiệu tiếng Việt. Tiếng Việt là một ngôn ngữ phức tạp với hệ thống dấu câu phong phú, tạo ra thách thức lớn trong việc nhận diện chính xác các ký tự. Các nghiên cứu trước đây chủ yếu tập trung vào bộ dữ liệu tiếng Anh và chưa giải quyết được những thách thức đặc thù của tiếng Việt. Chính vì thế, *Hierarchical Vietnamese Signboard Dataset* ra đời nhằm lấp đầy khoảng trống đó.

Cụ thể, nghiên cứu bao gồm: (1) Xây dựng *Hierarchical Vietnamese Signboard Dataset*, bộ dữ liệu bảng hiệu tiếng Việt phân cấp đầu tiên được chú thích ở ba cấp độ (từ, dòng và đoạn văn) và đa góc chụp: mỗi bảng hiệu được chụp từ ba góc (từ trái sang, chính diện và từ phải sang); (2) Đề xuất phương pháp mới và đánh giá hiệu quả của một số phương pháp SOTA trong phát hiện và phân tích bố cục văn bản trên bộ dữ liệu này. Không những thế, *Hierarchical Vietnamese Signboard Dataset* còn có thêm thông tin về tọa độ GPS, nhằm tạo ra một bộ dữ liệu chất lượng, đa dạng về góc chụp, có thể hữu ích cho nhiều nhiệm vụ khác.



Hình 1. Hợp nhất hai nhiệm vụ phát hiện và phân tích bố cục văn bản trên bảng hiệu tiếng Việt

Chúng tôi hy vọng rằng nghiên cứu này sẽ đóng góp vào việc phát triển các ứng dụng thực tiễn trong việc tự động hóa quá trình nhận diện và phân tích văn bản trong các tài liệu và hình ảnh thực tế, đặc biệt là trong ngữ cảnh tiếng Việt.

GIỚI THIỆU

Scene text detection là quá trình nhận diện và định vị văn bản xuất hiện trong các hình ảnh chụp từ môi trường thực tế [1]. Trong khi đó, document layout analysis là quá trình xác định và phân loại các thành phần khác nhau của một tài liệu, chẳng hạn như tiêu đề, đoạn văn, hình ảnh, bảng biểu và chú thích [2]. Với sự phát triển mạnh mẽ của công nghệ thị giác máy tính và học sâu, việc hợp nhất hai nhiệm vụ đó đã trở thành một nhánh nghiên cứu quan trọng. Trong quá trình nghiên cứu, chúng tôi nhận thấy bảng hiệu là một dạng thông tin có bố cục đặc biệt, hữu ích cho các ứng dụng thực tế như thu thập và phân tích dữ liệu trong bảng hiệu để nghiên cứu thị trường hay tự động cập nhật tên đường, cửa hàng và các địa điểm khác lên bản đồ số,... Việc phát hiện và phân tích bố cục trên bảng hiệu là một nhiệm vụ không chỉ đòi hỏi khả năng nhận diện chính xác các thành phần văn bản mà còn cần phải hiểu rõ mối quan hệ giữa các thành phần đó, tương tự như việc gom nhóm các từ có cùng ý nghĩa về mặt nội dung và ngữ nghĩa trong NLP.

Unified Detectors là mô hình đầu tiên hợp nhất hai nhiệm vụ này vào một mô hình duy nhất, sử dụng CNN để trích xuất đặc trưng (feature extraction) và Transformer để học các mối quan hệ không gian và ngữ nghĩa giữa các thành phần trong ảnh [3].

Dựa trên những ý tưởng đó, chúng tôi đề xuất một phương pháp mới để giải quyết bài toán trên là kết hợp kiến trúc *Transformer* và module *Interactive Attention* [7].

Phương pháp này cho phép mô hình học các mối quan hệ tương tác giữa các thành phần văn bản trong ảnh. Sự kết hợp này giúp tối ưu hóa quá trình học sâu, giúp mô hình học được các đặc trưng phức tạp từ dữ liệu ảnh một cách hiệu quả hơn.

Sau khi hoàn thiện mô hình, chúng tôi sẽ tiến hành huấn luyện trên bộ dữ liệu bảng hiệu tiếng Việt phân cấp do nhóm tự thu thập, được tham khảo từ bộ *HierText* [4][5]. Để đánh giá hiệu quả, chúng tôi sẽ so sánh mô hình của mình với các mô hình SOTA trước đó (*Unified Detector* [3], *Upstage KR*[6]), nhằm kiểm chứng khả năng cải thiện hiệu suất trong việc phát hiện và phân tích bối cảnh văn bản trên bảng hiệu tiếng Việt.

Mô tả bài toán:

- **Input:** Ảnh chứa một bảng hiệu.
- **Output:** Danh sách các đoạn văn bản phát hiện được trên bảng hiệu. Mỗi đoạn có nhãn để phân loại các thành phần trên bảng hiệu (tên cửa hàng, địa chỉ...) và danh sách các dòng văn bản. Mỗi dòng gồm danh sách các từ với tọa độ đỉnh bounding box và nội dung từ.



Hình 2. Ví dụ về kết quả của bài toán Phát hiện văn bản và Phân tích bối cảnh văn bản

MỤC TIÊU

- Xây dựng bộ dữ liệu bảng hiệu tiếng Việt phân cấp - Hierarchical Vietnamese Signboard, bộ dữ liệu được chú thích theo cấu trúc phân cấp (hierarchical annotations) để phục vụ cho quá trình huấn luyện và đánh giá.
- Nghiên cứu hợp nhất hai nhiệm vụ Scene Text Detection và Document Layout Analysis và đề xuất phương pháp mới.
- Đánh giá hiệu xuất của một số phương pháp SOTA và phương pháp do nhóm đề xuất trên bộ dữ liệu Hierarchical Vietnamese Signboard.

NỘI DUNG VÀ PHƯƠNG PHÁP

Trong đề tài này, chúng tôi sẽ thực hiện nghiên cứu các nội dung chính sau:

❖ **Nội dung 1:** Tìm hiểu tổng quan đề tài

➤ Phương pháp: tìm hiểu tổng quan các phương pháp để giải quyết từng nhiệm vụ scene text detection và document layout analysis.

❖ **Nội dung 2:** Nghiên cứu các phương pháp hợp nhất hai nhiệm vụ phát hiện và phân tích bối cảnh văn bản xuất hiện trên ảnh (scene text) hiện có.

➤ Phương pháp: tìm hiểu các phương pháp SOTA hiện có như Unified Detector, Upstage KR, . . . từ các công trình đã được công bố trên các top conference.

❖ **Nội dung 3:** Nghiên cứu và đề xuất phương pháp mới

➤ Phương pháp:

- Nghiên cứu kiến trúc của Transformer và Interactive Attention.
- Xây dựng model để giải quyết bài toán.

❖ **Nội dung 4:** Xây dựng bộ dữ liệu Hierarchical Vietnamese Signboard

➤ Phương pháp:

- Tạo guideline hướng dẫn để thu thập dữ liệu và lên kế hoạch label.
- Sử dụng điện thoại, máy ảnh để chụp bảng hiệu của các cửa hàng. Sau đó chọn lọc những bảng hiệu đạt yêu cầu và bắt đầu xử lý, gán nhãn.

❖ **Nội dung 5:** Chạy thực nghiệm và đánh giá trên bộ dữ liệu Hierarchical Vietnamese Signboard.

➤ Phương pháp:

- Chạy thực nghiệm các phương pháp hiện có và phương pháp do chúng tôi đề xuất trên bộ dữ liệu Hierarchical Vietnamese Signboard.
- Thống kê và đánh giá các phương pháp.

KẾT QUẢ MONG ĐỢI

- Bộ dữ liệu Hierarchical Vietnamese Signboard.
- Phương pháp mới để hợp nhất hai nhiệm vụ phát hiện và phân tích bố cục văn bản xuất hiện trong ảnh (scene text), cụ thể là trên bảng hiệu tiếng Việt.
- Kết quả đánh giá giữa phương pháp đề xuất và các phương pháp SOTA hiện có, trên bộ dữ liệu Hierarchical Vietnamese Signboard.

TÀI LIỆU THAM KHẢO

- [1]. Shangbang Long, Xin He, Cong Yao: Scene Text Detection and Recognition: The Deep Learning Era. Int. J. Comput. Vis. 129(1): 161-184 (2021)
- [2]. Jilin Wang, Michael Krumdick, Baojia Tong, Hamima Halim, Maxim Sokolov, Vadym Barda, Delphine Vendryes, Chris Tanner: A Graphical Approach to Document Layout Analysis. ICDAR (5) 2023: 53-69
- [3]. Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, Michalis Raptis: Towards End-to-End Unified Scene Text Detection and Layout Analysis. CVPR 2022: 1039-1049
- [4]. Maoyuan Ye, Jing Zhang, Juhua Liu, Chenyu Liu, Baocai Yin, Cong Liu, Bo Du, Dacheng Tao: Hi-SAM: Marrying Segment Anything Model for Hierarchical Text Segmentation. CoRR abs/2401.17904 (2024)
- [5]. Shangbang Long, Siyang Qin, Yasuhisa Fujii, Alessandro Bissacco, Michalis Raptis: Hierarchical Text Spotter for Joint Text Spotting and Layout Analysis. WACV 2024: 892-902
- [6]. Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, Michalis Raptis: ICDAR 2023 Competition on Hierarchical Text Detection and Recognition. ICDAR (2) 2023: 483-497
- [7]. Xingyu Wan, Chengquan Zhang, Pengyuan Lyu, Sen Fan, Zihan Ni, Kun Yao, Errui Ding, Jingdong Wang: Towards Unified Multi-granularity Text Detection with Interactive Attention. CoRR abs/2405.19765 (2024)