


THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
<https://youtu.be/oxjYVhvU35M>
- Link slides (dạng .pdf đặt trên Github của nhóm):
<https://github.com/VanKhaiiii/CS519.O11-STR>
- *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: Trương Văn Khải● MSSV: 21520274 	<ul style="list-style-type: none">● Lớp: CS519.O11● Tự đánh giá (điểm tổng kết môn): 9/10● Số buổi vắng: 0● Số câu hỏi QT cá nhân: 14/14● Link Github: https://github.com/VanKhaiiii/CS519.O11-STR● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none">○ Lên ý tưởng bài toán○ Viết đề cương, làm poster, làm slide○ Làm video YouTube
--	---

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

XỬ LÝ ẢNH NGOẠI CẢNH: NGHIÊN CỨU VIỆC PHÁT HIỆN VÀ NHẬN DIỆN VĂN BẢN TIẾNG VIỆT VỚI DBNETPP VÀ PARSEQ.

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

OUTDOOR IMAGE PROCESSING: INVESTIGATING TEXT DETECTION AND RECOGNITION IN VIETNAMESE USING DBNETPP AND PARSEQ.

TÓM TẮT (*Tối đa 400 từ*)

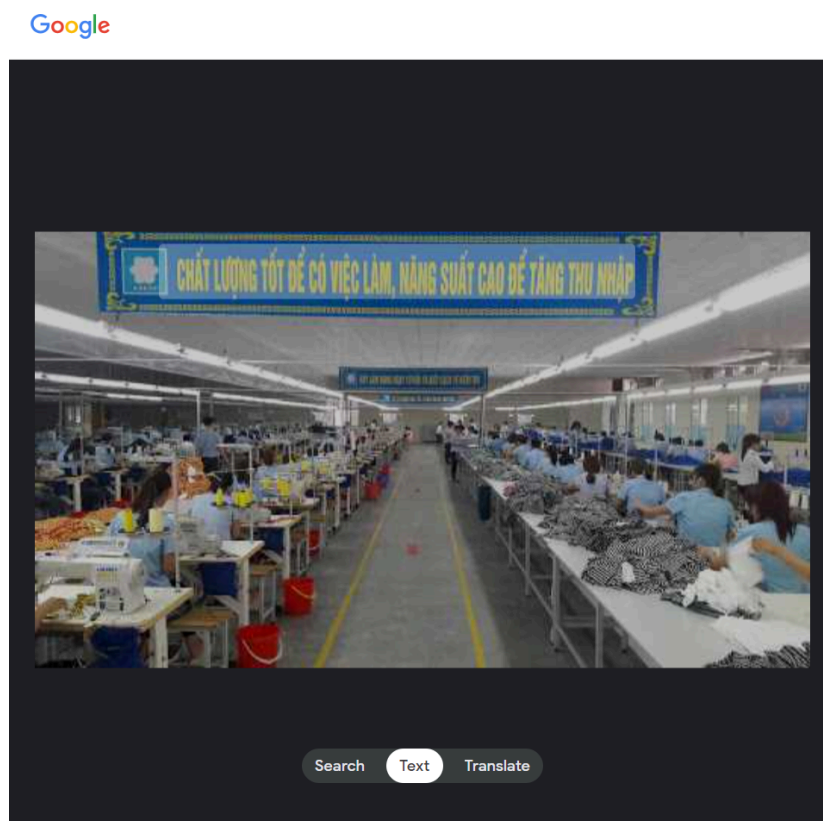
Chữ viết là một trong những phát minh vĩ đại nhất trong lịch sử loài người. Chữ viết đã tồn tại song song cùng con người qua hàng ngàn năm, nhờ có chữ viết mà các bài học đã lưu truyền lại được qua các thế hệ để con người ngày có thể học lại và ứng dụng nó. Vì vậy trong thời đại hiện đại hóa, làm sao để chúng ta có thể chuyển đổi các “bài học” đó để lưu trữ dưới dạng số hóa một cách tự động và nhanh hơn để phục vụ cho nhiều mục đích khác nhau? Đó cũng chính là lý do ra đời của rất nhiều công cụ chuyển đổi văn bản dưới dạng giấy, ảnh chụp sang dạng kỹ thuật số, có thể kể đến như: **Google Lens** [1], **Microsoft Seeing AI** [2], ... Đi kèm với sự tăng trưởng không ngừng của công nghệ học sâu, bài toán Phát hiện và Nhận diện văn bản trong cảnh, đặc biệt là văn bản tiếng Việt đã thu hút sự tìm tòi về nghiên cứu của tôi trong lĩnh vực này.

Trong đề tài này, tôi sẽ tập trung vào việc kết hợp các mô hình dựa trên hướng tiếp cận là tận dụng các mô hình SOTA ở các tác vụ riêng lẻ trong bài toán Phát hiện và Nhận diện văn bản nói chung. Thứ nhất, tôi sẽ đề xuất việc sử dụng mô hình **DBNetpp** [3] cho bài toán phát hiện văn bản tiếng Việt trong cảnh. Thứ hai, tôi đề xuất việc sử dụng đầu ra của **DBNetpp** [3] và đưa vào mô hình **PARSeq** [4] để phục vụ tác vụ nhận diện văn bản tiếng Việt. Cuối cùng, tôi sẽ xây dựng một chương trình ứng dụng để tăng tính thực tế cũng như kiểm chứng hiệu suất của việc sử dụng kết hợp hai mô hình này.

GIỚI THIỆU (*Tối đa 1 trang A4*)

Trong những năm gần đây, các phương pháp học sâu dần trở thành xu thế mới trong

hướng giải quyết bài toán Phát hiện và Nhận diện văn bản trong cảnh, có thể kể đến có phương pháp đã khá thành công trên dữ liệu tiếng Việt như: **ABCNet** [5], **Mask TextSpotter v3** [6] hay **Dictionary-guided Scene Text Recognition** [7] của **VinAI**. Tuy nhiên, các phương pháp này đều làm cho pipeline của bài toán trở nên phức tạp hơn hay việc phải sử dụng nguồn tài nguyên lớn về phần cứng lớn trong quá trình huấn luyện. Bên cạnh đó, theo hiểu biết của tôi, việc sử dụng các công cụ có sẵn để xử lý nhận diện văn bản tiếng Việt vẫn còn rất hạn chế, và chỉ có công ty Google đã công bố công nghệ. Tuy nhiên điểm hạn chế là Google bắt người dùng phải trả phí và họ không công khai công nghệ sử dụng.



Hình 1: Tính năng Text cho phép người dùng Phát hiện và Nhận diện các vùng văn bản có trong ảnh của Google.

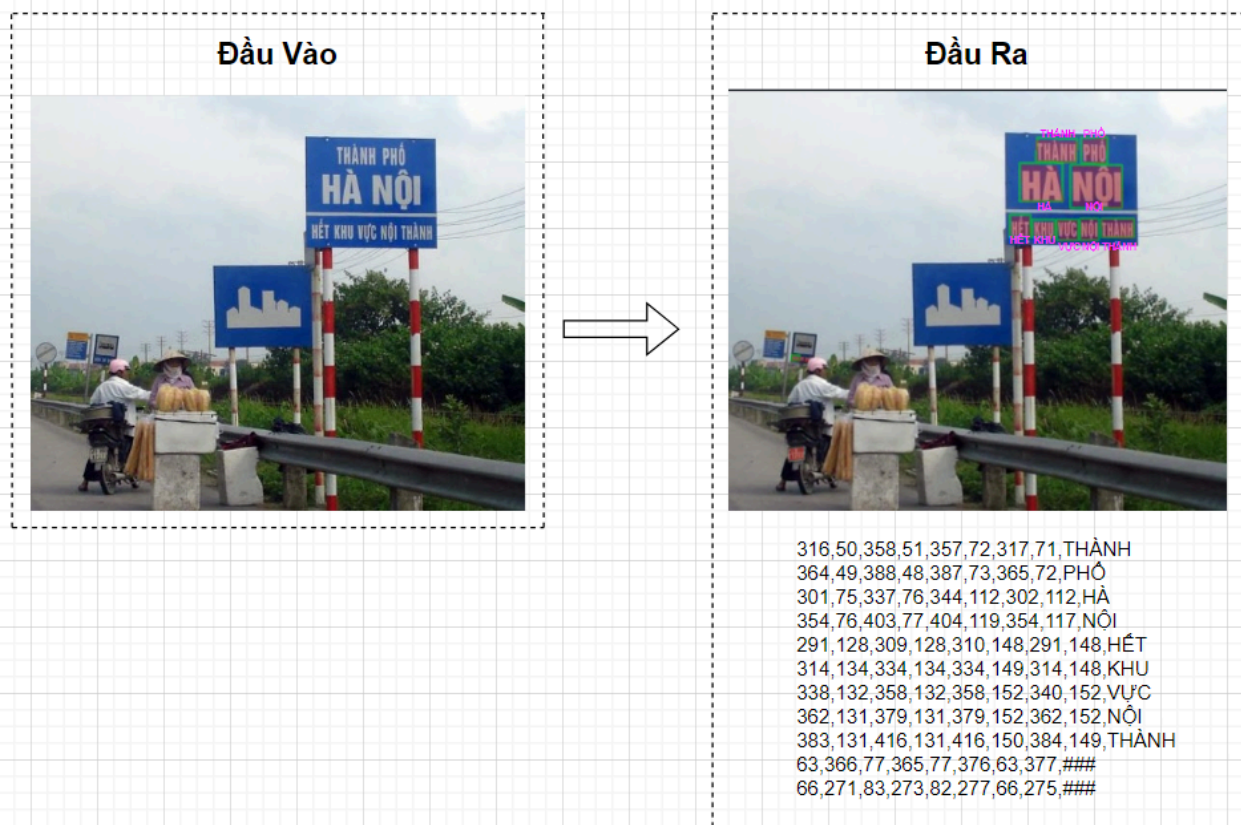
Bên cạnh việc các kiến trúc mô hình áp dụng trong bài toán này vẫn luôn có sự phát triển mạnh mẽ, song bài toán này vẫn đối diện với nhiều hạn chế như hình ảnh cảnh có khác nhau về kích thước, về phong chữ hay hình dạng của văn bản, về sự tác động của các yếu tố bên ngoài như: con người, thời tiết, bối cảnh phức tạp, ...

Các mô hình **DBNetpp** [3] hay **PARSeq** [4] đã có nhiều thành công nhất định trong từng tác

vụ riêng biệt của chúng. **DBNetpp** [3] có được độ chính xác cao và tốc độ nhanh nhất so với các phương pháp trước đó trên tác vụ Phát hiện văn bản. **PARSeq** [4] giúp cải thiện khả năng dự đoán, việc học từ theo nhiều hướng của **PARSeq** [4] có thể chú ý đến nhiều cách dự đoán từ khác nhau.

Trong đề tài này, tôi sẽ nghiên cứu thuật toán của mô hình **DBNetpp** [3] cho bài toán Phát hiện văn bản với đầu vào là ảnh cảnh. Sau đó, tôi sẽ cố gắng tiếp tục sử dụng đầu ra của giai đoạn trước để làm đầu vào cho mô hình **PARSeq** [4] - xương sống của giai đoạn Nhận diện văn bản, cuối cùng trả về các văn bản tiếng Việt tương ứng. Cụ thể, đầu vào và đầu ra của toàn bộ quy trình trên sẽ được phát biểu như sau:

- **Đầu vào:** Ảnh chứa nội dung văn bản bằng Tiếng Việt, với mỗi hình ảnh có thể chứa một hoặc nhiều đối tượng văn bản và với các biến dạng khác nhau như kích thước, màu sắc...
- **Đầu ra:** Các **polygons** xác định vùng chứa văn bản trong ảnh và chuỗi ký tự tương ứng. Mỗi **polygon** là một danh sách tọa độ (x, y) của các điểm.



Hình 2: Mô tả đầu vào và đầu ra của bài toán

MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

- Nghiên cứu thuật toán **DBNetpp** [3], **PARSeq** [4] hiện có, tìm hiểu cách kết hợp và áp dụng cài đặt thử nghiệm chúng trong việc Phát hiện và Nhận diện các văn bản tiếng Việt trong ảnh cảnh.
- Tiến hành đánh giá phương pháp kết hợp này trên một bộ dữ liệu mới. Phân tích các trường hợp đúng, sai của từng giai đoạn để rút được điểm mạnh, yếu của từng mô hình trên dữ liệu tiếng Việt.
- Xây dựng ứng dụng web trực quan hóa kết quả dự đoán.

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

- Nội dung:
 - Tìm hiểu các thuật toán **DBNetpp** [3], **PARSeq** [4] có thể tiếp nhận xử lý ảnh đầu vào với đa dạng kích thước, đa dạng hướng chụp (chụp ảnh ở xa hay gần, chụp chữ nghiêng, ngang, dọc, ...) và kết hợp chúng trong cùng một quy trình xử lý.
 - Tự xây dựng một bộ dữ liệu mới **UIT-STR20k** gồm 20.000 ảnh ngoại cảnh với độ đa dạng về kích thước mẫu dữ liệu, bối cảnh chụp, độ phân giải, để mô hình có thể học từ tập dữ liệu và đưa ra những dự đoán có tính minh bạch.
 - Tiến hành huấn luyện tuần tự mô hình **DBNetpp** [3], **PARSeq** [4] trên tập dữ liệu mới, sau đó tiến hành phân tích và đánh giá kết quả.
 - Tìm hiểu cách đánh giá hiệu suất riêng cho từng giai đoạn: **Precision**, **Recall**, **HmeanIOUMetric** cho giai đoạn Phát hiện văn bản. **Accuracy**, **OneMinusNED** cho giai đoạn Nhận diện văn bản và phương pháp đánh giá **End2End** cho toàn bộ các giai đoạn của quy trình.
 - Xây dựng chương trình ứng dụng minh họa.
- Phương pháp:
 - Tìm hiểu kiến trúc và nguyên lý hoạt động của hai mô hình **DBNetpp** [3] và **PARSeq** [4] chạy được với tất cả kích thước, hướng chụp ảnh đầu vào. Tìm

cách thiết lập các thông số để phù hợp, các phương pháp tăng cường dữ liệu nhằm đáp ứng nhu cầu về hiệu suất sử dụng của mô hình.

- Tôi tạo một bộ dữ liệu mới **UIT-STR20k** bằng cách tự thu thập các hình ảnh ngoại cảnh từ nhiều nguồn trên Internet có xuất hiện văn bản tiếng Việt về biển quảng cáo, biển báo, sách, áo có in chữ, ... hay cũng tự chụp ảnh bằng điện thoại về các phong cảnh đường phố với nhiều thiết bị điện thoại khác nhau, nhiều góc chụp khác nhau, ở các thời điểm khác nhau trong ngày để đảm bảo sự đa dạng về kích thước đầu vào của ảnh, bối cảnh, kiểu chữ hay phong chữ, ... Kết quả thu được sẽ là một bộ dữ liệu gồm 20.000 ảnh ngoại cảnh, với kích thước của từng mẫu dữ liệu là khác nhau, độ phân giải cũng khác nhau để tăng tính minh bạch của bộ dữ liệu. Quy trình đánh nhãn của bộ dữ liệu sẽ được tuân theo quy trình đánh nhãn của tập dữ liệu **VinText** [7], bộ dữ liệu dành riêng cho cuộc thi **AI Challenge 2021** cho bài toán Nhận diện chữ tiếng Việt trong ảnh ngoại cảnh.
- Tiến hành cài đặt thử nghiệm và huấn luyện tuần tự mô hình **DBNetpp** [3], thu được một tập ảnh là các vùng ảnh được **DBNetpp** [3] dự đoán là có xuất hiện văn bản. Tiếp đến sử dụng tập ảnh đầu ra của **DBNetpp** [3] làm đầu vào huấn luyện cho mô hình **PARSeq** [4]. Cả quá trình trên được chạy trên bộ dữ liệu đã tự thu thập, phân tích và đánh giá kết quả dựa trên các tiêu chí đánh giá tiêu chuẩn như **Precision**, **Recall**, **Hmean** cho giai đoạn Phát hiện văn bản; **Accuracy**, **OneMinusNED** cho giai đoạn Nhận diện văn bản và **FPS** cho tác vụ xử lý về tốc độ thực tế. Sau đó tiến hành đánh giá kết quả cuối cùng dựa theo phương pháp đánh giá **End2End**.
- Xây dựng ứng dụng trên nền web cho phép người dùng tải lên một hay nhiều ảnh để kiểm chứng, cho phép phóng to thu nhỏ các ảnh tải lên để tăng tính trải nghiệm của người dùng. Ngoài ra, người dùng có thể chọn tải về các kết quả dự đoán của toàn bộ quá trình nhằm phục vụ các mục đích cá nhân khác.

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

Sau đây sẽ là một số đóng góp dự kiến sẽ đạt được trong quá trình tôi thực hiện nghiên cứu:

- Xây dựng bộ dữ liệu gồm 20.000 ảnh ngoại cảnh, được đánh nhãn nghiêm ngặt, đảm bảo được độ đa dạng về kích thước từng mẫu dữ liệu, bối cảnh ảnh, kiểu chữ, kích thước chữ ... trên văn bản tiếng Việt. Dự kiến toàn bộ bộ dữ liệu sẽ bao gồm hơn 500.000 trường văn bản.
- Báo cáo các phương pháp kỹ thuật của các mô hình **DBNetpp** [3], **PARSeq** [4] được sử dụng trong bài toán Phát hiện và Nhận diện văn bản tiếng Việt trong ảnh ngoại cảnh. Kết quả thực nghiệm, phân tích và đánh giá phương pháp sử dụng.
- Chương trình minh họa trên nền web về Phát hiện và Nhận diện văn bản, tương tự như: [Google Lens](#) (Chỉ làm về phần Phát hiện và Nhận diện các văn bản có trong ảnh, không thực hiện thêm các tác vụ Dịch hay Tìm kiếm với thông tin được trích xuất).

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. [Google Lens](#)

[2]. <https://www.microsoft.com/en-us/garage/wall-of-fame/seeing-ai/>

[3]. Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, Xiang Bai, “Real-Time Scene Text Detection with Differentiable Binarization and Adaptive Scale Fusion,” IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), tập 45, số 1, pp. 919 - 931, 2022.

[4]. Darwin Bautista, Rowel Atienza, “Scene Text Recognition with Permuted Autoregressive Sequence Models,” trong the 17th European Conference on Computer Vision (ECCV 2022), Tel-Aviv, 2022.

[5]. Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, Liangwei Wang, “ABCNet: Real-Time Scene Text Spotting With Adaptive Bezier-Curve Network,” trong 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020.

[6]. Minghui Liao, Guan Pang, Jing Huang, Tal Hassner, Xiang Bai, “Mask TextSpotter v3: Segmentation Proposal Network for Robust Scene Text Spotting,” trong ECCV 2020, Glasgow, Scotland, 2020.

[7]. Nguyen Nguyen, Thu Nguyen, Vinh Tran, Minh-Triet Tran, Thanh Duc Ngo, Thien Huu Nguyen, Minh Hoai, “Dictionary-guided Scene Text Recognition,” trong 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville,

TN, USA, 2021.