THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút): https://www.youtube.com/watch?v=9Ds2L92Kt6w
- Link slides (dạng .pdf đặt trên Github): https://github.com/DyThen-Kumo/CS519.O21.KHTN/blob/main/slide/report.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 Duy Thắng • MSSV: 22521333 	• Lớp: CS519.O21.KHTN • Tự đánh giá (điểm tổng kết môn): 8.5/10 • Số buổi vắng: 0 • Số câu hỏi QT cá nhân: 10 • Link Github: https://github.com/DyThen-Kumo/CS519.O21.KHTN

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

MÔ HÌNH NHẬN DIỆN KÝ TỰ QUANG HỌC (OCR) SỬ DỤNG TRANSFORMER CHO NHẬN DIỆN CHỮ VIẾT TAY TIẾNG VIỆT.

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

OPTICAL CHARACTER RECOGNITION (OCR) MODEL USING TRANSFORMER FOR VIETNAMESE HANDWRITING RECOGNITION.

TÓM TẮT *(Tối đa 400 từ)*

Nhận diện chữ viết tay tiếng Việt gặp nhiều thách thức do tính đa dạng và phức tạp của các ký tự và dấu thanh. Các mô hình truyền thống thường kết hợp CNN để xử lý hình ảnh, RNN để tạo ra đầu ra và một mô hình ngôn ngữ để sửa lỗi chính tả, dẫn đến hệ thống phức tạp và khó tối ưu hóa. Kiến trúc Transformer, đã đạt nhiều thành công trong xử lý ngôn ngữ tự nhiên (NLP) và thị giác máy tính (CV), có thể giải quyết những hạn chế này. Tuy nhiên, các mô hình OCR dựa trên Transformer hiện nay thường được huấn luyện trên bộ dữ liệu tiếng Anh, giảm hiệu quả khi áp dụng cho tiếng Việt. Việc huấn luyện lại từ đầu rất tốn kém tài nguyên và thời gian. Liệu có cách nào để khắc phục điều đó? Câu trả lời là tinh chỉnh (finetuning) một mô hình đã được đào tạo trên ngôn ngữ có bảng chữ cái Latinh để đào tạo thêm về tiếng Việt vì mô hình đã học được một số nét chữ cơ bản. Nghiên cứu này nhằm finetuning mô hình OCR sử dụng kiến trúc Transformer (TrOCR) để nhận diện chữ viết tay tiếng Việt, mang lại hiệu suất cao hơn và khả năng ứng dụng rộng rãi hơn.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Optical Character Recognition (OCR) là công nghệ quan trọng cho việc chuyển đổi văn bản từ hình ảnh thành dạng chữ viết số, đóng vai trò then chốt trong nhiều ứng dụng như số hóa tài liệu, nhận dạng chữ viết tay, và trích xuất thông tin từ hình ảnh.

Một hệ thống OCR thường gồm hai phần chính: Text Detection và Text Recognition.

Về Text Detection, có thể sử dụng các mô hình hiệu suất cao như YOLO để phát hiện văn bản. Do đó, nghiên cứu này sẽ chủ yếu tập trung vào Text Recognition. Về Text Recognition, các mô hình trước đây thường theo dạng CRNN: sử dụng CNN để hiểu nội dung hình ảnh và RNN để sinh ra output. Quá trình này khiến việc huấn luyện khá lâu và mất nhiều công sức, tuy nhiên, nó cũng khá giống cấu trúc Encoder-Decoder của Transformer.

Trong vài năm gần đây, các mô hình Transformer đã chứng tỏ khả năng vượt trội trong nhiều tác vụ NLP và CV. Với khả năng xử lý mạnh mẽ, Transformer không chỉ phù hợp cho việc hiểu ngữ cảnh trong văn bản mà còn có thể áp dụng hiệu quả cho các nhiệm vụ OCR.

Quá trình fine-tune mô hình OCR sử dụng Transformer (TrOCR) cho dữ liệu chữ viết tay tiếng Việt là một bước quan trọng nhằm hiểu và xử lý tốt hơn những đặc trưng của dữ liệu chữ viết tay tiếng Việt, bao gồm các dấu thanh, dấu câu, các ký tự đặc biệt và cách viết của mỗi người, từ đó cải thiện độ chính xác của kết quả nhận dạng.

Input: Một hoặc nhiều hình ảnh chữ viết tay tiếng Việt và dataset chữ viết tay tiếng Việt đã được gán nhãn.

Output: Văn bản được trích xuất từ hình ảnh đầu vào. Văn bản này có thể được lưu dưới dạng chuỗi ký tự hoặc dưới dạng tệp văn bản (.txt) hoặc các định dạng tài liệu khác (ví dụ: .docx, .pdf).

MỤC TIÊU

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

1. Cải thiện độ chính xác **Word Accuracy** của mô hình TrOCR trong tác vụ nhận diện hình ảnh chữ viết tay tiếng Việt.

   $$Word\ Accuracy = \frac{Words\ Correct}{Words\ Correct + Words\ Misspelled + Words\ Skipped + Words\ Added}$$

2. Đào tạo một mô hình đào tạo sẵn (pre-trained model) để dễ dàng tùy chỉnh cho nhiều dữ liệu hình ảnh chữ viết tay tiếng Việt với đa dạng kiểu chữ hơn.

3. Áp dụng mô hình cho một ứng dụng cụ thể: nhận diện chữ viết tay tiếng Việt từ một hình ảnh đầu vào trong ứng dụng Translate.

NỘI DUNG VÀ PHƯƠNG PHÁP

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

1. NỘI DUNG

- Nghiên cứu các hệ thống OCR hiện nay và cách chúng hoạt động.
- Nghiên cứu về cấu trúc Transformer và quá trình xử lý của nó.
- Nghiên cứu về TrOCR, mô hình OCR theo Transformer đã được đào tạo sẵn cho các tác vụ nhận dạng chữ viết tay.
- Thu thập dữ liệu chữ viết tay tiếng Việt đã được gán nhãn để có bộ dataset đủ lớn để có thể finetuning mà không bị overfitting.
- Huấn luyện mô hình TrOCR base trên bộ dataset đã thu thập, đánh giá kết quả và lưu lại pre-trained model.
- Xây dựng ứng dụng cho pre-trained model đã có.

2. PHƯƠNG PHÁP

- **Thu Thập và Chuẩn Bị Dữ Liệu:** Dữ liệu chữ viết tay tiếng Việt đã được gắn nhãn được thu thập và chuẩn bị từ Internet. Dữ liệu này sau đó được chia thành các tập train, validation, test.
- **Mô Hình Huấn Luyện Ban Đầu:** Pre-trained TrOCR (TrOCR-Base-IAM), được huấn luyện trên một tập dữ liệu lớn và đa dạng (IAM Dataset) để học các đặc trưng cơ bản của hình ảnh chữ viết tay tiếng Anh, được lựa chọn để tiến hành finetuning.
- **FineTuning với Dữ Liệu Chữ Viết Tay Tiếng Việt:** Pre-trained model sẽ được finetuning trên tập dữ liệu chữ viết tay tiếng Việt. Quá trình này bao gồm việc điều chỉnh các tham số mô hình để phù hợp với các đặc điểm ngữ cảnh và hình ảnh của tiếng Việt.

- **Đánh Giá và Tinh Chỉnh:** Mô hình được đánh giá trên tập validation để kiểm tra hiệu suất. Dựa trên các kết quả này, mô hình sẽ được tinh chỉnh thêm để đạt được độ chính xác cao nhất rồi lưu lại làm pre-trained model.
- **Xây Dựng Ứng Dụng được tích hợp Pre-Trained Model:** Xây dựng một ứng dụng dịch máy đơn giản, cho phép upload một bức ảnh chữ viết tay tiếng Việt và nhận diện nội dung bức ảnh làm input cho phần dịch nghĩa sang ngôn ngữ khác.

KẾT QUẢ MONG ĐỢI

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

- **Độ chính xác của mô hình:** Độ chính xác mong đợi dựa theo Word Accuracy trung bình là lớn hơn 80%.
- **Pre-trained Model:** Mô hình đào tạo trước sẽ có kích thước ít hơn 1.5GB để dễ dàng cài đặt vô ứng dụng.
- **Ứng dụng được tích hợp Pre-trained Model:** Ứng dụng có thể cho phép người dùng upload hình ảnh chữ viết tay tiếng Việt và nhận diện chính xác.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Zhoujun Li, Furu Wei:
TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models. AAAI 2023: 13094-13102

[2]. Jan Kohút, Michal Hradis:
Fine Tuning Is a Surprisingly Effective Domain Adaptation Baseline in Handwriting Recognition. CoRR abs/2302.06308 (2023)

[3]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin:
Attention is All You Need. NIPS 2017: 5998-6008

[4]. Michael Jungo, Lars Vögtlin, Atefeh Fakhari, Nathan Wegmann, Rolf Ingold, Andreas Fischer, Anna Scius-Bertrand:
Impact of Ground Truth Quality on Handwriting Recognition. CoRR abs/2312.09037 (2023)

[5]. Vittorio Pippi, Silvia Cascianelli, Christopher Kermorvant, Rita Cucchiara:
How to Choose Pretrained Handwriting Recognition Models for Single Writer Fine-Tuning. CoRR abs/2305.02593 (2023)

[6]. Mst. Shapna Akter, Hossain Shahriar, Alfredo Cuzzocrea, Nova Ahmed, Carson K. Leung:
Handwritten Word Recognition using Deep Learning Approach: A Novel Way of Generating Handwritten Words. CoRR abs/2303.07514 (2023)