TẠP CHÍ KHOA HỌC VÀ CÔNG NGHỆ ĐẠI HỌC DUY TÂN
DTU Journal of Science and Technology 04(77) (2026) 75-86
ĐẠI HỌC DUY TÂN
DUYTAN UNIVERSITY

Đánh giá hiệu năng mô hình CNN-RNN-CTC và TrOCR trong nhận dạng chữ viết tay tiếng Việt từ hình ảnh tài liệu

Performance Evaluation of CNN-RNN-CTC and TrOCR in Vietnamese Handwritten Text Recognition from Document Images

Nguyễn Thị Thanh Phương[a], Ngô Văn Hiếu[a*], Phan Long[a], Phạm Phú Khương[a]
Nguyen Thi Thanh Phuong[a], Ngo Van Hieu[a*], Phan Long[a], Pham Phu Khuong[a]

Khoa Công nghệ Thông tin, Trường Khoa học Máy tính và Trí tuệ Nhân tạo, Đại học Duy Tân, Đà Nẵng, Việt Nam
Faculty of Information Technology, School of Computer Science and Artificial Intelligence, Duy Tan University, 55000, Danang, Vietnam

(Ngày nhận bài: 10/4/2026, ngày phản biện xong: 11/8/2026, ngày chấp nhận đăng: 18/8/2026)

Tóm tắt

Bài báo trình bày nghiên cứu so sánh mô hình học sâu lai CNN-RNN-CTC và mô hình TrOCR trong bài toán nhận dạng chữ viết tay tiếng Việt, được thực hiện trên hai tập dữ liệu HANDS-VN (VNOnDB2018) và Mì AI Handwritten Dataset với quy trình thực nghiệm thống nhất nhằm đảm bảo tính khách quan. Kết quả cho thấy mô hình TrOCR đạt hiệu năng tốt hơn khi Accuracy tăng từ 54,37% lên 78,92%, đồng thời CER giảm từ 45,63% xuống 29,85% và WER giảm từ 76,27% xuống 55,41% so với CNN-RNN-CTC. Kết quả này khẳng định ưu thế của kiến trúc Transformer trong việc khai thác ngữ cảnh toàn cục, từ đó giảm đáng kể lỗi ở cả mức ký tự và từ. Bên cạnh ưu thế về độ chính xác của TrOCR, sự đánh đổi giữa hiệu năng nhận dạng và chi phí tính toán của hai kiến trúc vẫn cần được đánh giá đầy đủ hơn trong các nghiên cứu tiếp theo.

Từ khoá: nhận dạng chữ viết tay tiếng Việt, OCR, CNN-RNN-CTC, TrOCR, Transformer

Abstract

This paper presents a comparative study of the hybrid deep learning model CNN-RNN-CTC and the TrOCR model for Vietnamese handwritten text recognition. The experiments are conducted on two datasets, HANDS-VN (VNOnDB2018) and the Mì AI Handwritten Dataset, using a unified experimental pipeline to ensure objectivity. The results show that TrOCR achieves better performance, with Accuracy increasing from 54.37% to 78.92%, while CER decreases from 45.63% to 29.85% and WER decreases from 76.27% to 55.41% compared with CNN-RNN-CTC. These results demonstrate the advantage of the Transformer architecture in capturing global context, thereby significantly reducing errors at both the character and word levels. Despite TrOCR's advantage in recognition accuracy, the trade-off between recognition performance and computational cost of the two architectures requires further comprehensive evaluation in future studies.

Keywords: Vietnamese handwritten text recognition, OCR, CNN-RNN-CTC, TrOCR, Transformer

[*]*Tác giả liên hệ*: Ngô Văn Hiếu
Email: ngovanhieu2@dtu.edu.vn

1. Giới thiệu

1.1. Đặt vấn đề

Trong thực tế, việc nhận dạng văn bản từ hình ảnh tài liệu, đặc biệt là chữ viết tay, vẫn còn nhiều thách thức và chưa đạt được độ chính xác cao trong nhiều trường hợp. Nguyên nhân chủ yếu đến từ sự đa dạng trong cách viết của con người. Chữ viết tay chịu ảnh hưởng bởi thói quen cá nhân, tốc độ viết, trình độ viết chữ, cũng như phong cách viết của từng người. Nhiều người có nét chữ khó đọc, ký tự bị biến dạng, viết tắt hoặc nối liền giữa các chữ, dẫn đến việc nhận dạng tự động trở nên phức tạp.

Đối với tiếng Việt, bài toán này càng trở nên khó khăn hơn do đặc thù ngôn ngữ có dấu. Các dấu thanh như sắc, huyền, hỏi, ngã, nặng thường được viết không rõ ràng, sai vị trí hoặc bị lược bỏ trong quá trình viết nhanh. Ngoài ra, sự khác biệt trong cách viết giữa các cá nhân, ví dụ như độ nghiêng, kích thước hoặc mức độ liền nét của chữ, làm tăng sự đa dạng của dữ liệu đầu vào. Điều này gây khó khăn lớn cho các hệ thống nhận dạng văn bản, đặc biệt khi dữ liệu không được chuẩn hóa.

Trong bối cảnh đó, việc nghiên cứu và đánh giá các mô hình học sâu nhằm cải thiện khả năng nhận dạng chữ viết tay tiếng Việt từ hình ảnh tài liệu, đặc biệt là trong điều kiện dữ liệu thực tế như chữ viết khó đọc hoặc không đồng nhất, là một vấn đề cần thiết.

1.2. Giải quyết vấn đề

Để giải quyết các thách thức nêu trên, bài báo này tập trung nghiên cứu và so sánh hai hướng tiếp cận tiêu biểu trong nhận dạng văn bản bằng học sâu.

Thứ nhất, mô hình lai Convolutional Neural Network [1] - Recurrent Neural Network [2] - Connectionist Temporal Classification [3] (CNN-RNN-CTC) là một kiến trúc truyền thống đã được sử dụng rộng rãi trong bài toán nhận dạng văn bản. Trong đó CNN được sử dụng để trích xuất đặc trưng từ hình ảnh đầu vào, RNN có nhiệm vụ mô hình hóa quan hệ chuỗi giữa các đặc trưng theo thời gian và CTC đóng vai trò ánh xạ chuỗi đặc trưng thành chuỗi ký tự mà không cần căn chỉnh thủ công giữa đầu vào và đầu ra. Nhờ sự kết hợp này, mô hình CNN-RNN-CTC có khả năng nhận dạng văn bản theo trình tự, đặc biệt phù hợp với dữ liệu chữ viết tay.

Thứ hai, mô hình TrOCR [4], đại diện cho hướng tiếp cận hiện đại dựa trên Transformer, sử dụng cơ chế attention để học trực tiếp ánh xạ từ hình ảnh sang chuỗi văn bản. Mô hình này có khả năng nắm bắt ngữ cảnh toàn cục tốt hơn và giảm phụ thuộc vào các bước xử lý trung gian.

Trong nghiên cứu này, hai mô hình được triển khai và huấn luyện trên cùng cùng bộ dữ liệu được kết hợp từ hai tập dữ liệu chữ viết tay tiếng Việt, với quy trình tiền xử lý và thiết lập thực nghiệm thống nhất. Hiệu năng của các mô hình được đánh giá thông qua các chỉ số Accuracy [5], Character Error Rate (CER) [6] và Word Error Rate (WER) [7] nhằm đảm bảo tính khách quan và công bằng trong so sánh.

1.3. Đóng góp của bài báo

Xuất phát từ các vấn đề nêu trên, bài báo này tập trung so sánh hiệu năng của mô hình học sâu lai CNN-RNN-CTC và mô hình TrOCR trong bài toán nhận dạng chữ viết tay tiếng Việt từ hình ảnh tài liệu. Cả hai mô hình đều được huấn luyện trên cùng hai tập dữ liệu, với quy trình tiền xử lý, thiết lập huấn luyện và tiêu chí đánh giá được thống nhất nhằm đảm bảo tính công bằng trong so sánh.

Các đóng góp chính của bài báo bao gồm: (i) xây dựng quy trình thực nghiệm thống nhất cho việc đánh giá hiệu năng các mô hình nhận dạng văn bản; (ii) thực hiện so sánh trực tiếp mô hình học sâu lai CNN-RNN-CTC và mô hình TrOCR trên cùng điều kiện dữ liệu và cấu hình siêu tham số, thông qua các chỉ số CER và WER; (iii) phân tích kết quả thực nghiệm để làm rõ ưu, nhược điểm của từng mô hình trong điều kiện dữ liệu

thực tế như chữ viết khó đọc, khác biệt phong cách viết giữa các cá nhân và dữ liệu không đồng nhất.

2. Các nghiên cứu liên quan

2.1. Nhận dạng văn bản từ hình ảnh tài liệu

Nhận dạng văn bản từ hình ảnh tài liệu là một bài toán quan trọng Optical Character Recognition (OCR) [8], là một bài toán quan trọng trong lĩnh vực xử lý ảnh và trí tuệ nhân tạo nhằm chuyển đổi nội dung văn bản trong ảnh sang dạng văn bản có thể xử lý được bởi máy tính. Trong đó, nhận dạng chữ viết tay được xem là thách thức hơn do sự đa dạng trong cách viết của từng cá nhân, sự biến dạng của ký tự, cũng như ảnh hưởng của các yếu tố như nhiễu, điều kiện ánh sáng và chất lượng hình ảnh.

Đối với tiếng Việt, bài toán nhận dạng văn bản càng trở nên phức tạp do đặc thù ngôn ngữ có dấu và sự đa dạng trong phong cách viết giữa các cá nhân. Các dấu thanh có thể bị viết sai vị trí hoặc bị lược bỏ, trong khi các ký tự thường bị dính nét hoặc biến dạng, gây khó khăn cho các hệ thống nhận dạng tự động. Do đó, việc nghiên cứu các phương pháp hiệu quả cho nhận dạng văn bản từ hình ảnh tài liệu tiếng Việt là một hướng đi cần thiết và có ý nghĩa thực tiễn cao.

2.2. Các phương pháp học sâu nền tảng trong phân tích văn bản

Mô hình dựa trên CNN-RNN-CTC là một trong những hướng tiếp cận nền tảng trong bài toán nhận dạng văn bản bằng học sâu. Alex Graves và các cộng sự (2006) [9] đề xuất phương pháp Connectionist Temporal Classification (CTC) cho phép mô hình học ánh xạ giữa chuỗi đầu vào và chuỗi đầu ra mà không cần căn chỉnh thủ công, tạo tiền đề cho các hệ thống nhận dạng văn bản hiện đại sau này.

Tiếp theo, sự phát triển của các mô hình học sâu trong xử lý ảnh đã thúc đẩy việc áp dụng CNN vào OCR. Đặc biệt, He và các cộng sự (2016) [10] đã đề xuất ResNet, giúp cải thiện đáng kể khả năng trích xuất đặc trưng sâu và được sử dụng rộng rãi làm mạng trích xuất đặc trưng trong các hệ thống nhận dạng văn bản.

Dựa trên các nền tảng này, Baoguang Shi và các cộng sự (2017) [11] đã đề xuất mô hình CRNN, kết hợp CNN, RNN và CTC để xây dựng một hệ thống nhận dạng văn bản đầu - cuối. Trong đó, CNN được sử dụng để trích xuất đặc trưng hình ảnh, RNN (thường là LSTM hoặc BiLSTM) dùng để mô hình hóa chuỗi, và CTC đảm nhiệm việc ánh xạ sang chuỗi ký tự.

Mặc dù đạt hiệu năng tốt trong nhiều bài toán, các nghiên cứu gần đây cho thấy các mô hình học sâu lai CNN-RNN-CTC vẫn tồn tại một số hạn chế. Việc phụ thuộc vào RNN làm giảm khả năng xử lý song song và hạn chế trong việc nắm bắt ngữ cảnh dài. Ngoài ra, cơ chế CTC có thể gây sai lệch trong các trường hợp dữ liệu phức tạp, đặc biệt với chữ viết tay có sự biến dạng lớn hoặc bố cục không chuẩn (khoảng cách ký tự không đều, dòng chữ lệch hoặc các ký tự dính nhau). Những hạn chế này đã thúc đẩy sự phát triển của các mô hình mới dựa trên Transformer trong các nghiên cứu gần đây.

2.3. TrOCR và các biến thể Transformer trong nhận dạng chữ viết tay

Sự phát triển của kiến trúc Transformer đã mở ra hướng tiếp cận mới cho bài toán nhận dạng văn bản từ hình ảnh. Tiêu biểu, Da Chang và các cộng sự (2024) [12] đã đề xuất phương pháp Low-Rank Adaptation (DLoRA)-TrOCR, trong đó áp dụng kỹ thuật LoRA để tối ưu hóa quá trình tinh chỉnh, giúp giảm đáng kể chi phí tính toán trong khi vẫn duy trì độ chính xác cao. Cùng năm, Li và các cộng sự (2024) [13] tiến hành nghiên cứu so sánh các mô hình Transformer trong nhận dạng chữ viết tay và chỉ ra rằng các mô hình như TrOCR có khả năng khai thác ngữ cảnh toàn cục hiệu quả thông qua cơ chế chú ý, từ đó giảm đáng kể tỷ lệ lỗi ký tự so với các phương pháp truyền thống.

Tiếp theo, Laziz Hamdi và các cộng sự (2025) [14] đề xuất PILOT, một kiến trúc OCR đầu - cuối có khả năng kết hợp nhận dạng văn bản và định vị không gian trong cùng một mô hình. PILOT sử dụng bộ mã hóa CNN nhẹ kết hợp với bộ giải mã Transformer để sinh đồng thời nội dung văn bản và các mã tọa độ, qua đó hỗ trợ nhận dạng toàn trang, nhận dạng theo vùng và định vị văn bản trong một kiến trúc thống nhất.

2.4. Nhận xét và khoảng trống nghiên cứu

Từ các nghiên cứu đã trình bày, có thể thấy các mô hình học sâu lai CNN-RNN-CTC là nền tảng quan trọng trong nhận dạng chữ viết tay, tuy nhiên còn hạn chế trong việc nắm bắt ngữ cảnh dài và phụ thuộc vào cơ chế CTC. Trong khi đó, các mô hình Transformer như TrOCR cho thấy hiệu năng vượt trội nhờ khả năng khai thác ngữ cảnh toàn cục thông qua cơ chế chú ý.

Mặc dù vậy, phần lớn các nghiên cứu hiện nay chủ yếu tập trung vào các bộ dữ liệu chuẩn hóa hoặc các ngôn ngữ phổ biến, trong khi các nghiên cứu trên dữ liệu chữ viết tay tiếng Việt vẫn còn hạn chế. Đặc biệt, các yếu tố thực tế như chữ viết khó đọc, sự khác biệt sự khác biệt về phong cách viết giữa các cá nhân và dữ liệu không đồng nhất chưa được đánh giá đầy đủ.

Ngoài ra, các nghiên cứu so sánh trực tiếp giữa CNN-RNN-CTC và TrOCR trong cùng điều kiện thực nghiệm, với cùng tập dữ liệu và cấu hình siêu tham số, vẫn còn chưa nhiều. Do đó, bài báo này tập trung thực hiện so sánh công bằng giữa hai mô hình nhằm làm rõ hiệu năng trong bối cảnh dữ liệu tiếng Việt.

3. Phương pháp nghiên cứu

Phần này trình bày phương pháp nghiên cứu được sử dụng để đánh giá hiệu năng của mô hình học sâu lai CNN-RNN-CTC và mô hình TrOCR trong bài toán nhận dạng chữ viết tay tiếng Việt từ hình ảnh tài liệu. Nghiên cứu được thực hiện theo hướng thực nghiệm, với quy trình thống nhất nhằm đảm bảo tính khách quan và công bằng trong so sánh.

3.1. Tổng quan phương pháp

Trong nghiên cứu này, phương pháp tiếp cận được xây dựng theo hướng thực nghiệm so sánh, trong đó mô hình lai CNN-RNN-CTC và mô hình TrOCR được triển khai, huấn luyện và đánh giá độc lập trên cùng cùng bộ dữ liệu được kết hợp từ hai tập dữ liệu chữ viết tay tiếng Việt nhằm đảm bảo tính khách quan. Dữ liệu chữ viết tay tiếng Việt được thu thập và tiền xử lý thông qua các bước chuẩn hóa kích thước ảnh, loại bỏ nhiễu và chuẩn hóa nhãn văn bản.

Sau đó, hai mô hình được huấn luyện cùng cách chia dữ liệu, quy trình đánh giá và các tiêu chí đánh giá. Mô hình CNN-RNN-CTC khai thác khả năng trích xuất đặc trưng và xử lý chuỗi, trong khi TrOCR sử dụng kiến trúc Transformer để học trực tiếp ánh xạ từ hình ảnh sang văn bản. Cuối cùng, hiệu năng của các mô hình được đánh giá thông qua các chỉ số Accuracy, CER và WER nhằm so sánh và phân tích kết quả một cách khách quan.

3.2. Dữ liệu và tiền xử lý

Trong nghiên cứu này, hai tập dữ liệu chữ viết tay tiếng Việt được sử dụng bao gồm HANDS-VN (VNOnDB2018) và Mì AI Handwritten Dataset, nhằm phản ánh đặc điểm thực tế của bài toán nhận dạng chữ viết tay Tiếng Việt.

Tập dữ liệu HANDS-VN (VNOnDB2018) được sử dụng trong cuộc thi ICFHR 2018, bao gồm khoảng 1.146 đoạn văn bản tiếng Việt, tương ứng với 7.296 dòng và hơn 480.000 ký tự, được thu thập từ 200 người viết khác nhau. Dữ liệu thể hiện sự đa dạng về phong cách viết, bao gồm các trường hợp chữ viết rõ ràng, chữ viết khó đọc, ký tự dính nét và sai lệch dấu.

Tập dữ liệu Mì AI Handwritten Dataset bao gồm 1.838 ảnh dòng chữ viết tay tiếng Việt, trong đó mỗi ảnh được lưu dưới định dạng .png, kèm theo nhãn (ground truth) ở định dạng .json.

Dữ liệu được thu thập từ nhiều người viết khác nhau, với sự đa dạng về nét chữ, kích thước, độ nghiêng và mật độ ký tự, phản ánh các điều kiện thực tế như ảnh chụp từ thiết bị di động hoặc văn bản viết nhanh.

Sau khi kết hợp hai tập dữ liệu, tổng cộng 9.134 ảnh/dòng văn bản được sử dụng trong nghiên cứu, bao gồm 7.296 dòng từ HANDS-VN và 1.838 ảnh dòng văn bản từ Mì AI Handwritten Dataset.

Việc lựa chọn kết hợp hai tập dữ liệu này nhằm tận dụng ưu điểm bổ trợ lẫn nhau. Cụ thể, HANDS-VN cung cấp dữ liệu chuẩn hóa với chất lượng cao và được kiểm soát tốt, trong khi Mì AI phản ánh các điều kiện thực tế với nhiều nhiễu và biến thể trong chữ viết. Sự kết hợp này giúp mô hình được huấn luyện trên cả dữ liệu chuẩn và dữ liệu thực tế, từ đó nâng cao khả năng tổng quát hóa và đánh giá chính xác hơn hiệu năng trong các tình huống ứng dụng thực tế.

Trước khi đưa vào huấn luyện, dữ liệu được tiền xử lý nhằm đảm bảo tính đồng nhất và nâng cao chất lượng đầu vào. Cụ thể, các ảnh được chuẩn hóa về kích thước, chuyển đổi về định dạng phù hợp và chuẩn hóa giá trị pixel. Đồng thời, các ảnh bị nhiễu hoặc lệch được xử lý nhằm cải thiện độ rõ nét. Đối với nhãn văn bản, dữ liệu được chuẩn hóa theo định dạng Unicode để đảm bảo tính nhất quán, đặc biệt đối với các ký tự có dấu trong tiếng Việt.

3.3. Gán nhãn và chia tập dữ liệu

Trong nghiên cứu này, dữ liệu được sử dụng là các tập dữ liệu đã được gán nhãn sẵn từ các nguồn ban đầu, do đó không thực hiện bước gán nhãn lại. Mỗi ảnh trong tập dữ liệu đều đã được ánh xạ với một chuỗi văn bản tương ứng (ground truth), phục vụ trực tiếp cho các phương pháp học có giám sát trong bài toán nhận dạng chữ viết tay Tiếng Việt.

Sau khi hoàn tất quá trình tiền xử lý và chuẩn hóa dữ liệu, toàn bộ dữ liệu từ hai tập được gộp lại nhằm tăng tính đa dạng và khả năng tổng quát hóa cho mô hình. Dữ liệu sau đó được phân chia thành ba tập con bao gồm tập huấn luyện, tập kiểm tra và tập kiểm thử theo tỷ lệ lần lượt là 70%, 15% và 15%. Việc phân chia được thực hiện một cách ngẫu nhiên nhưng có kiểm soát nhằm đảm bảo sự phân bố đồng đều về độ dài chuỗi, phong cách viết và mức độ khó của dữ liệu giữa các tập.

Để đảm bảo tính khách quan trong quá trình so sánh, cùng một cách phân chia dữ liệu được áp dụng cho cả hai mô hình CNN-RNN-CTC và TrOCR. Điều này giúp loại bỏ các sai lệch có thể phát sinh do khác biệt trong dữ liệu huấn luyện và kiểm thử, từ đó đảm bảo rằng sự khác biệt về hiệu năng giữa các mô hình chủ yếu đến từ kiến trúc và phương pháp học.

Ngoài ra, tập kiểm tra được sử dụng trong quá trình huấn luyện nhằm điều chỉnh siêu tham số và theo dõi quá trình hội tụ của mô hình, trong khi tập kiểm thử được giữ độc lập và chỉ sử dụng để đánh giá hiệu năng cuối cùng. Cách tiếp cận này giúp đảm bảo tính khách quan và độ tin cậy của kết quả thực nghiệm.

3.4. Mô hình học sâu lai CNN-RNN-CTC

Mô hình học sâu lai CNN-RNN-CTC được sử dụng trong nghiên cứu này nhằm học trực tiếp ánh xạ từ ảnh văn bản đầu vào sang chuỗi ký tự đầu ra theo hướng đầu - cuối. Kiến trúc mô hình gồm ba thành phần chính là CNN, RNN và CTC. Trong đó, CNN có nhiệm vụ trích xuất đặc trưng từ ảnh, RNN dùng để mô hình hóa quan hệ tuần tự trong chuỗi đặc trưng, còn CTC thực hiện ánh xạ chuỗi đặc trưng sang chuỗi nhãn mà không cần căn chỉnh thủ công giữa đầu vào và đầu ra.

Cụ thể, khối CNN gồm 7 lớp tích chập với số lượng bộ lọc lần lượt là 64, 128, 256, 256, 512, 512 và 512. Sáu lớp đầu sử dụng kernel kích thước 3×3, trong khi lớp cuối sử dụng kernel 2×2. Hàm kích hoạt ReLU được áp dụng sau các lớp tích chập, kết hợp với các lớp max-pooling để giảm kích thước không gian của đặc trưng

trước khi chuyển sang khối BiLSTM. Khối BiLSTM gồm 2 tầng xếp chồng, mỗi tầng sử dụng 256 hidden units cho mỗi hướng. Đầu ra của hai hướng thuận và ngược được kết hợp để tạo biểu diễn ngữ cảnh hai chiều trước khi đưa qua lớp Softmax và cơ chế CTC để sinh chuỗi ký tự đầu ra.

Trước hết, với ảnh đầu vào I mô hình CNN được sử dụng để trích xuất các đặc trưng không gian của văn bản, chẳng hạn như nét chữ, biên ký tự, hình dạng và cấu trúc cục bộ của từng vùng ảnh. Quá trình này được biểu diễn như sau:

$$F = CNN(I) \quad (1)$$

Trong đó I là ảnh đầu vào F là bản đồ đặc trưng thu được sau các lớp tích chập và gộp mẫu. Công thức (1) cho thấy CNN đóng vai trò chuyển đổi ảnh gốc thành một biểu diễn đặc trưng có mức trừu tượng cao hơn, giúp mô hình tập trung vào các thông tin quan trọng phục vụ nhận dạng.

Sau khi thu được bản đồ đặc trưng F, mô hình tiến hành biến đổi bản đồ này thành chuỗi đặc trưng theo chiều ngang của ảnh. Việc chuyển đổi này được thực hiện vì văn bản thường được đọc theo thứ tự từ trái sang phải, do đó các đặc trưng theo chiều ngang có thể được xem như một chuỗi theo thời gian:

$$X = \{x_1, x_2, \ldots, x_T\} \quad (2)$$

Trong đó, X là chuỗi đặc trưng đầu vào cho RNN, x_t là vector đặc trưng tại thời điểm t, và T là độ dài chuỗi đặc trưng. Công thức (2) cho thấy đầu ra của CNN không còn được xử lý như một ảnh hai chiều, mà được biểu diễn lại thành một chuỗi một chiều để thuận lợi cho việc học ngữ cảnh tuần tự.

Tiếp theo, chuỗi đặc trưng $X = \{x_1, x_2, \ldots, x_T\}$ được đưa vào mạng RNN để mô hình hóa mối quan hệ ngữ cảnh theo cả hai hướng trái sang phải và phải sang trái. Điều này đặc biệt quan trọng đối với bài toán nhận dạng văn bản, vì việc dự đoán một ký tự không chỉ phụ thuộc vào thông tin tại vị trí hiện tại mà còn chịu ảnh hưởng bởi các ký tự lân cận trong chuỗi.

Trong nghiên cứu này, mạng Bidirectional Long Short-Term Memory (BiLSTM) được sử dụng để khai thác thông tin ngữ cảnh hai chiều. Cụ thể, trạng thái ẩn tại thời điểm t được tính như sau:

$$\vec{h}_t = LSTM(x_t, \vec{h}_{t-1}) \quad (3)$$

$$\overleftarrow{h}_t = LSTM(x_t, \overleftarrow{h}_{t+1}) \quad (4)$$

$$h_t = [\vec{h}_t; \overleftarrow{h}_t] \quad (5)$$

Trong đó, $\vec{h}_t$ và $\overleftarrow{h}_t$ lần lượt là trạng thái ẩn theo hướng thuận và hướng ngược, còn h_t là vector đặc trưng cuối cùng tại thời điểm t sau khi kết hợp hai hướng. Nhờ đó, mỗi h_t chứa thông tin ngữ cảnh của toàn bộ chuỗi, giúp cải thiện khả năng nhận dạng ký tự trong các trường hợp chữ viết không rõ ràng hoặc bị biến dạng.

Từ chuỗi trạng thái ẩn $H = \{h_1, h_2, \ldots, h_T\}$, mô hình tiến hành dự đoán phân bố xác suất trên tập ký tự tại mỗi thời điểm thông qua lớp Softmax:

$$y_t = Softmax(Wh_t + b) \quad (6)$$

Trong đó, y_t à vector xác suất của các ký tự tại thời điểm t, W, b là tham số của mô hình.

Để ánh xạ chuỗi dự đoán sang chuỗi ký tự đầu ra mà không cần căn chỉnh thủ công, mô hình sử dụng lớp CTC. Xác suất của một chuỗi nhãn y được xác định bởi tổng xác suất của tất cả các đường đi π có thể ánh xạ về y:

$$P(y \mid x) = \sum_{\pi \in B^{-1}(y)} P(\pi \mid x) \quad (7)$$

Trong đó, B là hàm loại bỏ các ký tự lặp và ký tự trống, còn $B^{-1}(y)$ là tập các đường đi tương ứng với chuỗi nhãn y.

Xác suất của một đường đi π được tính như sau:

$$P(\pi \mid x) = \prod_{t=1}^{T} y_t^{\pi_t} \quad (8)$$

Cuối cùng, hàm mất mát CTC được sử dụng để huấn luyện mô hình:

$$L_{CTC} = -\log P(y \mid x) \quad (9)$$

Hàm mất mát này cho phép mô hình học trực tiếp từ dữ liệu mà không cần thông tin căn chỉnh giữa ảnh và nhãn, giúp đơn giản hóa quá trình huấn luyện và nâng cao hiệu quả nhận dạng.

Như vậy, toàn bộ quy trình của mô hình CNN-RNN-CTC có thể được hiểu như một chuỗi các bước xử lý liên tiếp. Trước hết, ảnh đầu vào được đưa qua mạng CNN để trích xuất các đặc trưng không gian quan trọng, giúp biểu diễn văn bản dưới dạng các đặc trưng có mức trừu tượng cao. Tiếp theo, các đặc trưng này được chuyển đổi thành chuỗi và đưa vào mạng BiLSTM để mô hình hóa mối quan hệ ngữ cảnh theo cả hai hướng, từ đó tăng khả năng nhận dạng ký tự trong bối cảnh chuỗi văn bản.

Sau đó, lớp Softmax được sử dụng để sinh ra phân bố xác suất của các ký tự tại từng thời điểm, làm cơ sở cho việc dự đoán chuỗi đầu ra. Cuối cùng, cơ chế CTC được áp dụng để ánh xạ chuỗi xác suất này thành chuỗi ký tự hoàn chỉnh mà không cần căn chỉnh thủ công giữa ảnh và nhãn. Nhờ đó, mô hình có thể học trực tiếp từ dữ liệu đầu vào và nhãn tương ứng một cách hiệu quả.

Cách tiếp cận này cho phép mô hình CNN-RNN-CTC xử lý tốt bài toán nhận dạng văn bản, đặc biệt trong các trường hợp dữ liệu phức tạp như chữ viết tay với nhiều biến thể về hình dạng và cấu trúc ký tự.

3.5. Mô hình TrOCR

Mô hình TrOCR được sử dụng trong nghiên cứu này là một kiến trúc dựa trên Transformer theo hướng encoder–decoder, cho phép học trực tiếp ánh xạ từ ảnh đầu vào sang chuỗi văn bản đầu ra. Khác với mô hình lai CNN-RNN-CTC, TrOCR không sử dụng RNN hay CTC mà dựa hoàn toàn vào cơ chế attention để mô hình hóa mối quan hệ giữa các thành phần trong dữ liệu.

Trong nghiên cứu này, mô hình TrOCR-base-handwritten được sử dụng và fine-tuning trên dữ liệu chữ viết tay tiếng Việt. Bộ mã hóa sử dụng ViT gồm 12 lớp Transformer, kích thước ẩn 768 và 12 attention heads; ảnh đầu vào được chuẩn hóa về kích thước 384 × 384 pixel và chia thành các patch 16 × 16. Bộ giải mã gồm 12 lớp Transformer, kích thước ẩn 1024, 16 attention heads và kích thước lớp feed-forward là 4096. Cơ chế cross-attention được sử dụng để liên kết đặc trưng ảnh từ encoder với quá trình sinh chuỗi văn bản.

Trước hết, ảnh đầu vào I được chia thành các patch và đưa vào encoder để trích xuất đặc trưng, được biểu diễn như sau:

$$Z = Encoder(I) \quad (10)$$

Trong đó, $Z = \{z_1, z_2, \ldots, z_T\}$ là chuỗi đặc trưng đầu ra của encoder. Trong mỗi lớp encoder, cơ chế self-attention được sử dụng để học mối quan hệ giữa các vùng trong ảnh:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (11)$$

Trong đó, Q, K, V lần lượt là truy vấn, khoá và giá trị, còn d_k là kích thước vector. Nhờ cơ chế này, mô hình có khả năng khai thác ngữ cảnh toàn cục thay vì chỉ các đặc trưng cục bộ như CNN.

Tiếp theo, decoder được sử dụng để sinh chuỗi ký tự đầu ra từng bước. Tại thời điểm t, xác suất của ký tự được tính như sau:

$$y_t = \text{Softmax}\left(\text{Decoder}(y_{<t}, Z)\right) \quad (12)$$

Trong đó, $y_{<t}$ là chuỗi ký tự đã sinh trước đó. Quá trình huấn luyện mô hình được thực hiện thông qua hàm mất mát cross-entropy:

$$L = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, Z) \quad (13)$$

Trong đó, y_t là ký tự thực tại thời điểm t. Hàm mất mát này cho phép mô hình học cách sinh chuỗi văn bản một cách tuần tự.

Đối với bài toán nhận dạng văn bản tiếng Việt, việc xử lý dữ liệu đầu vào và đầu ra cần được điều chỉnh phù hợp với đặc điểm ngôn ngữ. Cụ thể, các chuỗi văn bản được chuẩn hóa theo định dạng Unicode nhằm đảm bảo tính nhất quán của các ký tự có dấu. Đồng thời, quá trình tokenization được thực hiện ở mức ký tự hoặc subword để giữ nguyên cấu trúc của các ký tự tiếng Việt như "ă", "â", "ô", "ư", tránh việc tách rời dấu khỏi ký tự gốc. Ngoài ra, dữ liệu cũng bao gồm các trường hợp chữ viết tay không rõ ràng, sai dấu hoặc biến dạng ký tự, giúp mô hình học được các đặc trưng thực tế và nâng cao khả năng tổng quát hóa.

Như vậy, mô hình TrOCR thực hiện nhận dạng văn bản thông qua việc trích xuất đặc trưng toàn cục bằng encoder, kết hợp với decoder để sinh chuỗi ký tự đầu ra theo ngữ cảnh. So với mô hình CNN-RNN-CTC, TrOCR có khả năng khai thác ngữ cảnh tốt hơn, đặc biệt trong các trường hợp dữ liệu phức tạp như chữ viết tay tiếng Việt.

4. Thực nghiệm và kết quả

Phần này trình bày thiết lập thực nghiệm và các tiêu chí đánh giá nhằm so sánh hiệu năng của mô hình học sâu lai CNN-RNN-CTC và mô hình TrOCR trong bài toán nhận dạng văn bản từ hình ảnh tài liệu. Các thí nghiệm được thiết kế theo hướng đảm bảo tính nhất quán và công bằng, trong đó cả hai mô hình được huấn luyện và đánh giá trên cùng tập dữ liệu, cùng quy trình tiền xử lý và cùng tiêu chí đánh giá.

4.1. Thiết lập thực nghiệm

Do mô hình học sâu lai CNN-RNN-CTC và mô hình TrOCR đều thuộc nhóm mô hình học sâu với số lượng tham số lớn, quá trình huấn luyện đòi hỏi tài nguyên tính toán cao, đặc biệt là khả năng xử lý song song trên GPU. Trong nghiên cứu này, các thí nghiệm được triển khai trên nền tảng Google Colaboratory Pro+, sử dụng GPU NVIDIA Tesla T4 (16 GB VRAM), nhằm đảm bảo khả năng huấn luyện và tinh chỉnh mô hình trong thời gian hợp lý.

Mô hình CNN-RNN-CTC được huấn luyện từ đầu với các tham số khởi tạo ngẫu nhiên, trong khi TrOCR được khởi tạo từ mô hình tiền huấn luyện TrOCR-base-handwritten và tiếp tục fine-tuning trên dữ liệu chữ viết tay tiếng Việt.

Toàn bộ dữ liệu được gộp từ hai tập HANDS-VN và Mì AI, sau đó được phân chia thành ba tập con bao gồm tập huấn luyện, tập kiểm tra và tập kiểm thử theo tỷ lệ lần lượt là 70%, 15% và 15%. Tỷ lệ này được lựa chọn nhằm đảm bảo lượng dữ liệu cho quá trình huấn luyện, đồng thời vẫn giữ một phần dữ liệu độc lập để đánh giá khả năng tổng quát hóa của mô hình. Việc phân chia được thực hiện thống nhất cho cả hai mô hình nhằm đảm bảo tính công bằng trong quá trình so sánh.

4.2. Cấu hình siêu tham số

Trong nghiên cứu này, mô hình học sâu lai CNN-RNN-CTC và mô hình TrOCR được huấn luyện với cấu hình siêu tham số được thiết lập nhằm đảm bảo khả năng hội tụ hiệu quả trong điều kiện tài nguyên tính toán hạn chế, đồng thời vẫn duy trì tính công bằng trong quá trình so sánh.

Cụ thể, số epoch huấn luyện được thiết lập là 100 nhằm cho phép mô hình học đủ lâu để khai thác đặc trưng từ dữ liệu. Batch size được lựa chọn là 128 nhằm tận dụng khả năng xử lý song song của GPU, đồng thời giúp ổn định quá trình cập nhật trọng số. Bộ tối ưu Adam được sử dụng do khả năng hội tụ nhanh và ổn định trong các bài toán học sâu, với learning rate được thiết lập ở mức phù hợp để tránh dao động trong quá trình huấn luyện.

Bên cạnh đó, cơ chế dừng sớm với tham số patience = 20 được áp dụng nhằm ngăn chặn hiện tượng quá khớp. Khi hiệu năng trên tập

kiểm tra không còn cải thiện sau một số epoch nhất định, quá trình huấn luyện sẽ được dừng lại, giúp tối ưu thời gian huấn luyện và nâng cao khả năng tổng quát hóa của mô hình.

Đối với hàm mất mát, mô hình học sâu lai CNN-RNN-CTC sử dụng CTC Loss, trong khi mô hình TrOCR sử dụng hàm mất mát Cross-Entropy, phù hợp với đặc điểm của từng kiến trúc. Việc thiết lập các siêu tham số theo cùng một nguyên tắc giúp đảm bảo rằng sự khác biệt về hiệu năng chủ yếu đến từ mô hình thay vì cấu hình huấn luyện.

4.3. Bộ chỉ số đánh giá

Để đánh giá hiệu năng của các mô hình trong bài toán nhận dạng văn bản, nghiên cứu sử dụng ba chỉ số chính bao gồm Accuracy, CER, WER. Các chỉ số này được lựa chọn nhằm phản ánh đầy đủ hiệu năng của mô hình từ cấp độ toàn chuỗi đến cấp độ ký tự.

Accuracy được sử dụng để đo lường tỷ lệ số mẫu được nhận dạng hoàn toàn chính xác trên tổng số mẫu kiểm thử. Chỉ số này được định nghĩa như sau:

$$Accuracy = \frac{N_{correct}}{N_{total}} \quad (14)$$

Trong đó, $N_{correct}$ là số mẫu được dự đoán đúng hoàn toàn và N_{total} là tổng số mẫu.

Tuy nhiên, trong các bài toán nhận dạng văn bản, việc đánh giá ở cấp độ ký tự đóng vai trò quan trọng, đặc biệt đối với các ngôn ngữ có dấu như tiếng Việt. Do đó, tỷ lệ lỗi ký tự CER được sử dụng để đo mức độ sai lệch giữa chuỗi dự đoán và chuỗi thực:

$$CER = \frac{S + D + I}{N} \quad (15)$$

Trong đó, S, D, I lần lượt là số ký tự thay thế, xóa và chèn thêm, còn N là tổng số ký tự của chuỗi gốc. CER càng nhỏ cho thấy mô hình càng nhận dạng chính xác ở mức ký tự.

Tương tự, tỷ lệ lỗi từ WER được sử dụng để đánh giá ở cấp độ từ:

$$WER = \frac{S + D + I}{N} \quad (16)$$

Trong đó, các thành phần được tính trên đơn vị từ thay vì ký tự. WER phản ánh khả năng nhận dạng chính xác các từ hoàn chỉnh, từ đó đánh giá gián tiếp khả năng hiểu ngữ cảnh của mô hình.

Nhìn chung, việc kết hợp cả ba chỉ số giúp đánh giá mô hình một cách toàn diện. Trong đó, Accuracy phản ánh độ chính xác tổng thể, còn CER và WER cung cấp cái nhìn chi tiết hơn về các lỗi phát sinh trong quá trình nhận dạng.

4.4. Kết quả định lượng

Kết quả huấn luyện của mô hình học sâu lai CNN-RNN-CTC và mô hình TrOCR được tổng hợp trong Bảng 1, trong đó trình bày các chỉ số Accuracy, CER và WER trên tập kiểm thử.

Bảng 1. Kết quả đánh giá hiệu năng của các mô hình

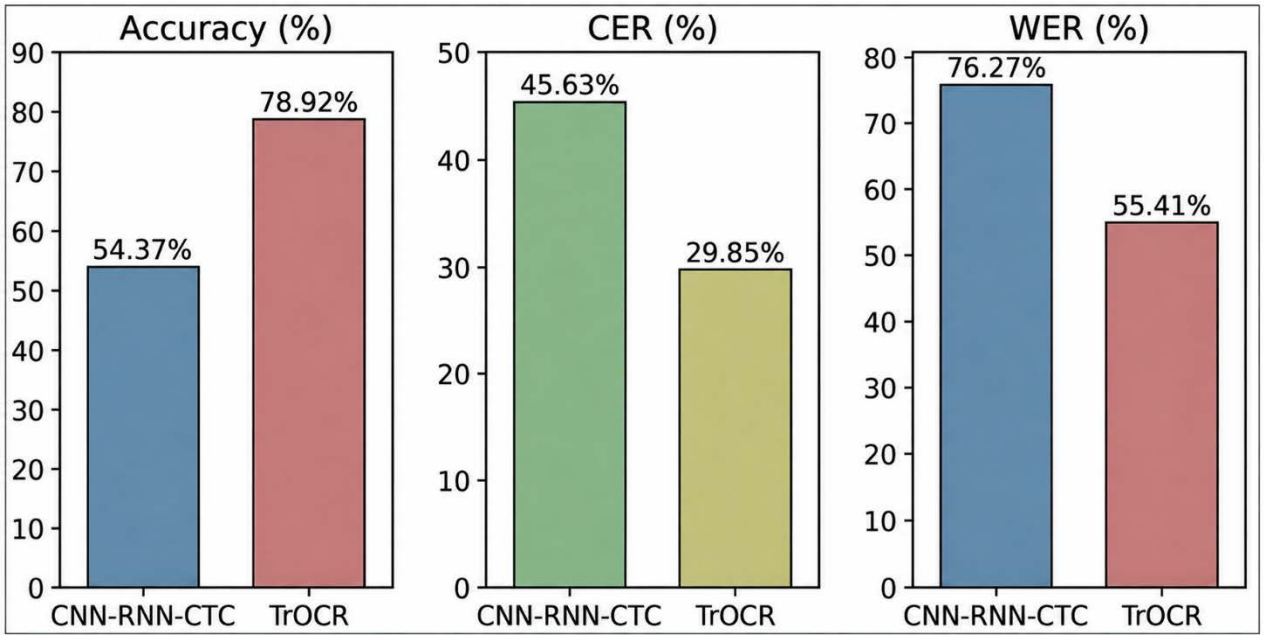
Mô hình	Accuracy (%)	CER (%)	WER (%)
CNN-RNN-CTC	54,37	45,63	76,27
TrOCR	78,92	29,85	55,41

Từ kết quả trong Bảng 1 có thể nhận thấy rằng mô hình học sâu lai CNN-RNN-CTC đạt mức hiệu năng cơ bản, tuy nhiên tỷ lệ lỗi ở cả cấp độ ký tự và từ vẫn còn cao, cho thấy khả năng nhận dạng còn hạn chế khi xử lý dữ liệu chữ viết tay phức tạp. Ngược lại, mô hình TrOCR đạt hiệu năng tốt hơn trên cả ba chỉ số, đặc biệt là giảm đáng kể CER và WER, cho thấy khả năng nhận dạng ổn định và chính xác hơn.

Bên cạnh đó, sự chênh lệch giữa CER và WER của hai mô hình cũng phản ánh mức độ tích lũy lỗi ở cấp độ từ, trong đó mô hình học sâu

lai CNN-RNN-CTC dễ bị lan truyền lỗi khi dự đoán sai ký tự. Trong khi đó, mô hình TrOCR cho thấy khả năng kiểm soát lỗi tốt hơn nhờ khai thác ngữ cảnh toàn cục trong quá trình dự đoán.

Để trực quan hóa kết quả trong Bảng 1, Hình 1 thể hiện các chỉ số đánh giá dưới dạng biểu đồ, giúp so sánh sự khác biệt về hiệu năng giữa hai mô hình một cách rõ ràng hơn.



Hình 1. Biểu đồ so sánh hiệu năng của hai mô hình CNN-RNN-CTC và TrOCR

5. Thảo luận

Kết quả thực nghiệm cho thấy sự khác biệt rõ rệt giữa mô hình học sâu lai CNN-RNN-CTC và mô hình TrOCR trong bài toán nhận dạng văn bản từ hình ảnh tài liệu. Sự khác biệt này không chỉ thể hiện ở các chỉ số định lượng mà còn phản ánh bản chất kiến trúc của từng mô hình.

Trước hết, mô hình học sâu lai CNN-RNN-CTC hoạt động dựa trên việc trích xuất đặc trưng cục bộ thông qua CNN và mô hình hóa chuỗi bằng RNN. Mặc dù cách tiếp cận này đã được chứng minh là hiệu quả trong nhiều bài toán OCR, nhưng vẫn tồn tại hạn chế trong việc khai thác ngữ cảnh dài hạn. Điều này dẫn đến việc mô hình dễ mắc lỗi trong các trường hợp ký tự bị biến dạng, dính nét hoặc khi chuỗi văn bản có cấu trúc phức tạp. Đối với chữ viết tay tiếng Việt, khó khăn này càng rõ hơn khi các dấu thanh hoặc dấu phụ được viết nhỏ, lệch vị trí hoặc gần với nét của các ký tự lân cận, làm cho thông tin thị giác tại một vùng riêng lẻ có thể chưa đủ để nhận dạng chính xác.

Ngược lại, mô hình TrOCR dựa trên kiến trúc Transformer có khả năng khai thác ngữ cảnh toàn cục thông qua cơ chế chú ý. Cơ chế tự chú ý cho phép mô hình học mối quan hệ giữa các vùng đặc trưng khác nhau của ảnh, trong khi bộ giải mã dự đoán ký tự dựa trên cả đặc trưng ảnh và chuỗi ký tự đã được sinh trước đó. Nhờ đó, mô hình có thể kết hợp thông tin thị giác với ngữ cảnh chuỗi trong quá trình nhận dạng, thay vì chỉ phụ thuộc vào đặc trưng tại một vị trí riêng lẻ. Cơ chế này đặc biệt hữu ích đối với chữ viết tay tiếng Việt, nơi dấu thanh, dấu phụ hoặc nét chữ có thể nhỏ, lệch vị trí và khó phân biệt. Đối với các ký tự có hình dạng gần nhau như “a”, “ă”, “â” hoặc “o”, “ô”, “ơ”, việc khai thác ngữ cảnh rộng hơn có thể hỗ trợ mô hình lựa chọn ký tự phù hợp hơn, từ đó góp phần giảm lỗi ở cấp độ ký tự và từ.

Một điểm đáng chú ý là sự khác biệt về chỉ số CER và WER giữa hai mô hình. Mặc dù cả hai mô hình đều có thể mắc lỗi ở cấp độ ký tự, nhưng mô hình học sâu lai CNN-RNN-CTC có xu

hướng tích lũy lỗi dẫn đến sai lệch lớn ở cấp độ từ. Trong khi đó, mô hình TrOCR có khả năng kiểm soát lỗi tốt hơn nhờ tận dụng thông tin ngữ cảnh, từ đó giảm thiểu tác động của các lỗi cục bộ. Cụ thể, CER giảm từ 45,63% xuống 29,85%, trong khi WER giảm từ 76,27% xuống 55,41%. Kết quả này cho thấy khả năng khai thác ngữ cảnh của TrOCR có thể góp phần giảm lỗi ở cấp độ ký tự và hạn chế sự tích lũy lỗi khi hình thành từ hoàn chỉnh.

Tuy nhiên, mô hình TrOCR cũng tồn tại một số hạn chế. Do nghiên cứu hiện tại chưa đánh giá thời gian huấn luyện, thời gian suy luận, số lượng tham số hoặc mức sử dụng bộ nhớ GPU, chưa đủ cơ sở thực nghiệm để kết luận mô hình nào có chi phí tính toán thấp hơn. Ngoài ra, hiệu năng của mô hình phụ thuộc nhiều vào chất lượng dữ liệu huấn luyện, đặc biệt trong các trường hợp dữ liệu không đồng nhất hoặc thiếu đa dạng.

Từ các phân tích trên, có thể thấy rằng TrOCR đạt hiệu năng nhận dạng tốt hơn CNN-RNN-CTC trên dữ liệu được khảo sát, đặc biệt trong các trường hợp chữ viết tay tiếng Việt có nhiều biến thể. Việc đánh giá thêm các chỉ số về chi phí tính toán sẽ được xem xét trong các nghiên cứu tiếp theo.

6. Kết luận và hướng phát triển

Bài báo đã trình bày và thực hiện so sánh hiệu năng của mô hình học sâu lai CNN-RNN-CTC và mô hình TrOCR trong bài toán nhận dạng văn bản từ hình ảnh tài liệu tiếng Việt. Các thí nghiệm được thiết lập trong cùng điều kiện dữ liệu, cấu hình và tiêu chí đánh giá nhằm đảm bảo tính công bằng và khách quan.

Kết quả thực nghiệm cho thấy mô hình TrOCR đạt hiệu năng vượt trội hơn so với mô hình học sâu lai CNN-RNN-CTC trên cả ba chỉ số Accuracy, CER và WER. Điều này khẳng định ưu thế của kiến trúc Transformer trong việc khai thác ngữ cảnh toàn cục và xử lý hiệu quả các chuỗi văn bản phức tạp, đặc biệt trong điều kiện dữ liệu chữ viết tay tiếng Việt có nhiều biến thể về nét chữ và dấu.

Tuy nhiên, nghiên cứu vẫn còn một số hạn chế nhất định, bao gồm quy mô dữ liệu chưa lớn và điều kiện thực nghiệm còn phụ thuộc vào tài nguyên tính toán. Trong tương lai, nghiên cứu có thể được mở rộng theo một số hướng chính. Thứ nhất, xây dựng và sử dụng các tập dữ liệu chữ viết tay tiếng Việt có quy mô lớn và đa dạng hơn nhằm nâng cao khả năng tổng quát hóa của mô hình. Thứ hai, áp dụng các kỹ thuật tăng cường dữ liệu) và tiền huấn luyện để cải thiện hiệu năng trong điều kiện dữ liệu hạn chế. Thứ ba, nghiên cứu kết hợp các mô hình lai giữa CNN và Transformer nhằm tận dụng ưu điểm của cả hai hướng tiếp cận. Cuối cùng, triển khai và đánh giá mô hình trong các hệ thống thực tế như số hóa tài liệu hoặc nhận dạng văn bản từ ảnh chụp nhằm kiểm chứng khả năng ứng dụng trong môi trường thực tế.

Tài liệu tham khảo

[1] Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2017). “ImageNet classification with deep convolutional neural networks”. *Communications of the ACM* (60(6)), 84–90. DOI: 10.1145/3065386.

[2] Lipton, Z.C., Berkowitz, J., & Elkan, C. (2015). *A critical review of recurrent neural networks for sequence learning*. Truy cập 12/8/2026, từ https://arxiv.org/abs/1506.00019.

[3] Graves, A. (2012). *Supervised sequence labelling with recurrent neural networks*. Berlin, Heidelberg: Springer.

[4] Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., & Wei, F. (2023). TrOCR: Transformer-based optical character recognition with pre-trained models. *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence, Washington, DC, USA, 2023* (tr. 13094–13102). Palo Alto, CA: AAAI Press.

[5] Powers, D.M.W. (2011). “Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation”. *Journal of Machine Learning Technologies* (2(1)), 37–63.

[6] Neudecker, C., Baierer, K., Gerber, M., Clausner, C., Antonacopoulos, A., & Pletschacher, S. (2021). A survey of OCR evaluation tools and metrics. *Proceedings of the 6th International Workshop on*

Historical Document Imaging and Processing (HIP '21), Lausanne, Switzerland, 2021 (tr. 13–18). New York, NY: Association for Computing Machinery.

[7] Morris, A.C., Maier, V., & Green, P. (2004). From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. *Proceedings of Interspeech 2004, Jeju Island, Korea, 2004* (tr. 2765–2768).

[8] Wang, X.F., He, Z.H., Wang, K., Wang, Y.F., Zou, L., & Wu, Z.Z. (2023). "A survey of text detection and recognition algorithms based on deep learning technology". *Neurocomputing* (556), 126702. DOI: 10.1016/j.neucom.2023.126702.

[9] Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, USA, 2006* (tr. 369–376). New York, NY: Association for Computing Machinery.

[10] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016* (tr. 770–778). Los Alamitos, CA: IEEE Computer Society.

[11] Shi, B., Bai, X., & Yao, C. (2017). "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition". *IEEE Transactions on Pattern Analysis and Machine Intelligence* (39(11)), 2298–2304. DOI: 10.1109/TPAMI.2016.2646371.

[12] Chang, D., & Li, Y. (2024). *DLoRA-TrOCR: Mixed Text Mode Optical Character Recognition Based On Transformer*. Truy cập 12/8/2026, từ https://arxiv.org/abs/2404.12734.

[13] Li, Y., Chen, D., Tang, T., & Shen, X. (2025). "HTR-VT: Handwritten text recognition with vision transformer". *Pattern Recognition* (158), 110967. DOI: 10.1016/j.patcog.2024.110967.

[14] Hamdi, L., Tamasna, A., Boisson, P., & Paquet, T. (2025). *VISTA-OCR: Towards generative and interactive end to end OCR models*. Truy cập 12/8/2026, từ https://arxiv.org/abs/2504.03621.