**NHẬN DẠNG CHỮ VIẾT TAY TIẾNG VIỆT   
SỬ DỤNG MÔ HÌNH TRANSFORMER VÀ ATTENTION OCR**

# Võ Ngọc Tín, Phạm Quang Huy

Khoa Công nghệ Thông tin, Trường Đại học Khoa học Tự nhiên – ĐHQG HCM

*19120685@student.hcmus.edu.vn, 19120535@student.hcmus.edu.vn*

***TÓM TẮT****:* *Bài báo này trình bày phương pháp nhận dạng chữ viết tay tiếng Việt kết hợp sử dụng mô hình mạng nơrơn tích chập (CNN) và Transformer. Mô hình kết hợp này có tính tổng quát tốt và có độ chính xác khá cao trên một bộ dataset mới dù chưa qua huấn luyện. Kết quả thử nghiệm cho thấy phương pháp đã đạt đến kết quả vượt trội so với phương pháp mô hình CRNN sử dụng CTCLoss.*

***Từ khóa:*** *Nhận dạng chữ viết tay tiếng Việt, mạng nơrơn tích chập CNN, mô hình Transformer OCR, mô hình Attention OCR, OCR.*

# GIỚI THIỆU

Nhận dạng chữ viết tay là một trong những ứng dụng quan trọng của ngành thị giác máy tính để giải quyết những vấn đề thực tiễn. Các nghiên cứu về nhận dạng chữ viết tay được thực hiện ở các ngôn ngữ như tiếng Anh, tiếng Nga, tiếng Trung, … và tiếng Việt. Mặc dù đã có một số nghiên cứu về nhận dạng chữ viết tay tiếng Việt, song vẫn còn chưa nhiều. Tiếng Việt là một ngôn ngữ đa dạng với 29 chữ cái và 6 thanh điệu, tuy nhiên bộ dữ liệu tiếng Việt lại tương đối ít. Bên cạnh đó, hiện nay mạng nơrơn học sâu đang nhận được sự quan tâm của đông đảo nhà nghiên cứu và được áp dụng hiệu quả trong rất nhiều lĩnh vực trong đó có xử lý chữ viết. Vì vậy, trong bài báo này sẽ giải quyết bài toán nhận dạng chữ viết tiếng Việt thông qua sử dụng một kiến trúc mạng nơrơn học sâu phổ biến là mạng nơrơn tích chập (Convolutional Neural Network - CNN [1]) kết hợp với Transformer (mô hình nền tảng của BERT [2]) thông qua thực nghiệm của 2 mô hình nhận dạng ký tự quang học (Optical Character Recognition – OCR [3]) là Attention OCR và Transformer OCR, trên tập dữ liệu có gần 10.000 ảnh chữ viết tay để đánh giá.

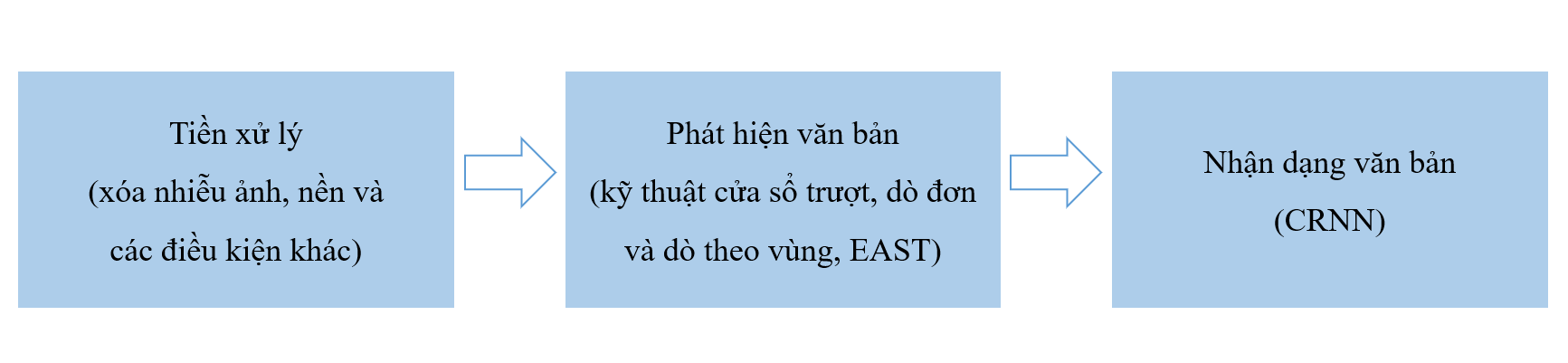
Attention OCR sử dụng kiến trúc attention sequence-to-sequence khá phổ biến trong các bài toán xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) và OCR. Việc kết hợp mô hìnhTransformer OCR với kiến trúc Transformer cho ra kết quả tỷ lệ nhận dạng đúng đạt 87,01 %, cho thấy sự tiến bộ vượt bậc so với phương pháp truyền thống là mô hình CRNN sử dụng CTCLoss.

Bài báo này được cấu trúc như sau: Các phương pháp đề xuất cho nhận dạng chữ viết tay Việt Nam được mô tả trong Phần II. Phần III sẽ là thực nghiệm các mô hình. Cuối cùng, chúng tôi sẽ đưa ra kết luận của mình và một số công việc cần xem xét cho tương lai trong Phần IV.

# PHƯƠNG PHÁP ĐỀ XUẤT

## Mạng nơron tích chập CNN của mô hình OCR

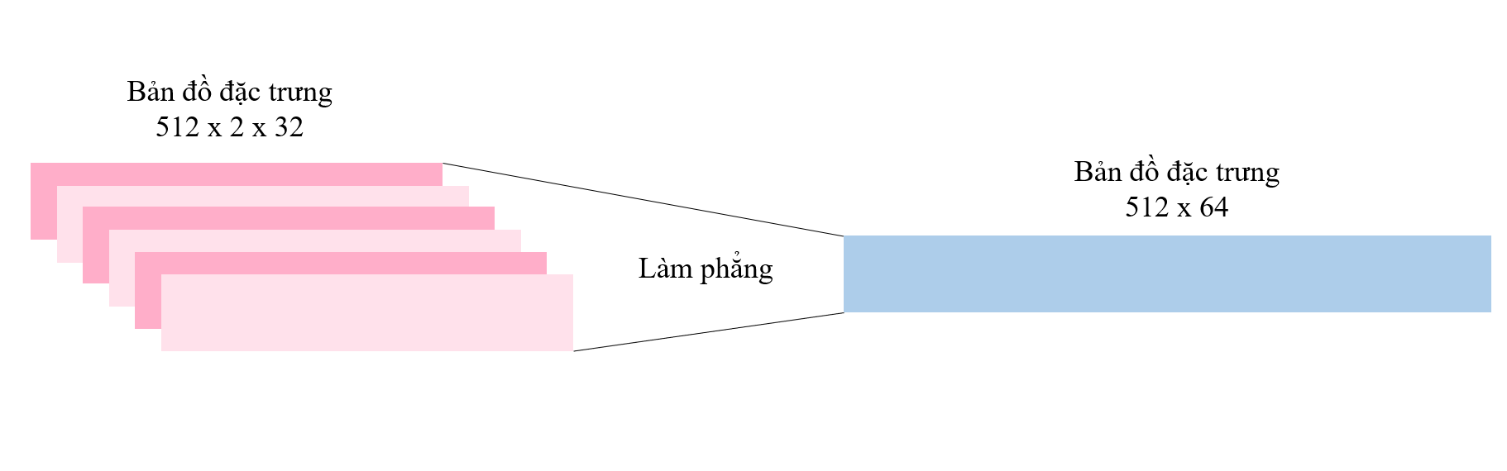
Hệ thống OCR thực hiện tuần tự theo các bước như trong hình 1. Đầu tiên, ảnh được đưa vào chế độ tiền xử lý, để loại bỏ nhiễu ảnh, xóa các nền phức tạp và xử lý các điều kiện ảnh khác nhau. Sau đó, kỹ thuật phát hiện văn bản tạo ra các giới hạn hộp xung quanh phần hình ảnh có văn bản. Để có thể phát hiện được văn bản, có thể dùng các kĩ thuật như cửa sổ trượt, dò đơn và dò theo vùng hoặc sử dụng trình phát hiện văn bản trong cảnh chính xác hiệu quả (EAST) [4]. Khi đã giới hạn được các văn bản, sử dụng kỹ thuật mạng thần kinh tái phát liên tục (CRNN) cho các tác vụ nhận dạng chuỗi dựa trên hình ảnh, như nhận dạng văn bản cảnh và OCR.



**Hình 1**. Quy trình thực thiện của hệ thống nhận dạng ký tự quang học

## Mô hình Attention OCR

Mô hình Attention OCR là sự kết hợp giữa mô hình mạng tích chập (CNN) và mô hình tập trung chuỗi sang chuỗi (attention sequence-to-sequence). Cách hoạt động của mô hình này tương tự như kiến trúc của mô hình chuỗi sang chuỗi trong bài toán dịch máy. Một ảnh qua mô hình mạng tích chập sẽ cho ra một bản đồ đặc trưng có kích thước kênh x chiều cao x chiều dài và bản đồ này sẽ trở thành đầu vào của mô hình bộ nhớ dài-ngắn hạn (LSTM). Cuối cùng chiều cao x chiều dài của bản đồ đặc trưng sẽ được duỗi thẳng, lúc này bản đồ có kích thước phù hợp với yêu cầu của mô hình LSTM.

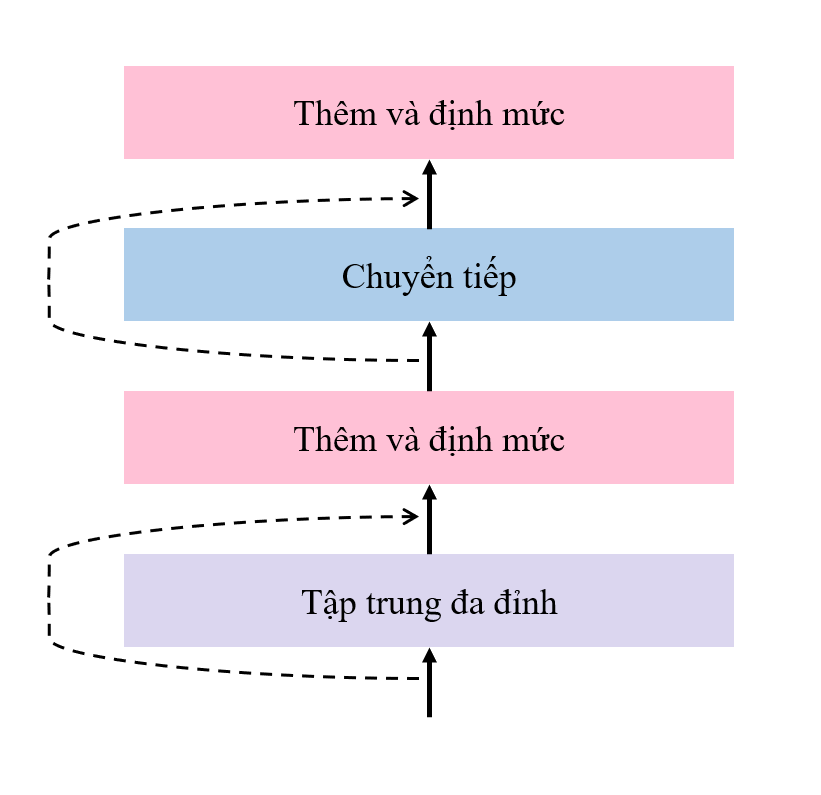


**Hình 2**. Mô phỏng mô hình Attention OCR

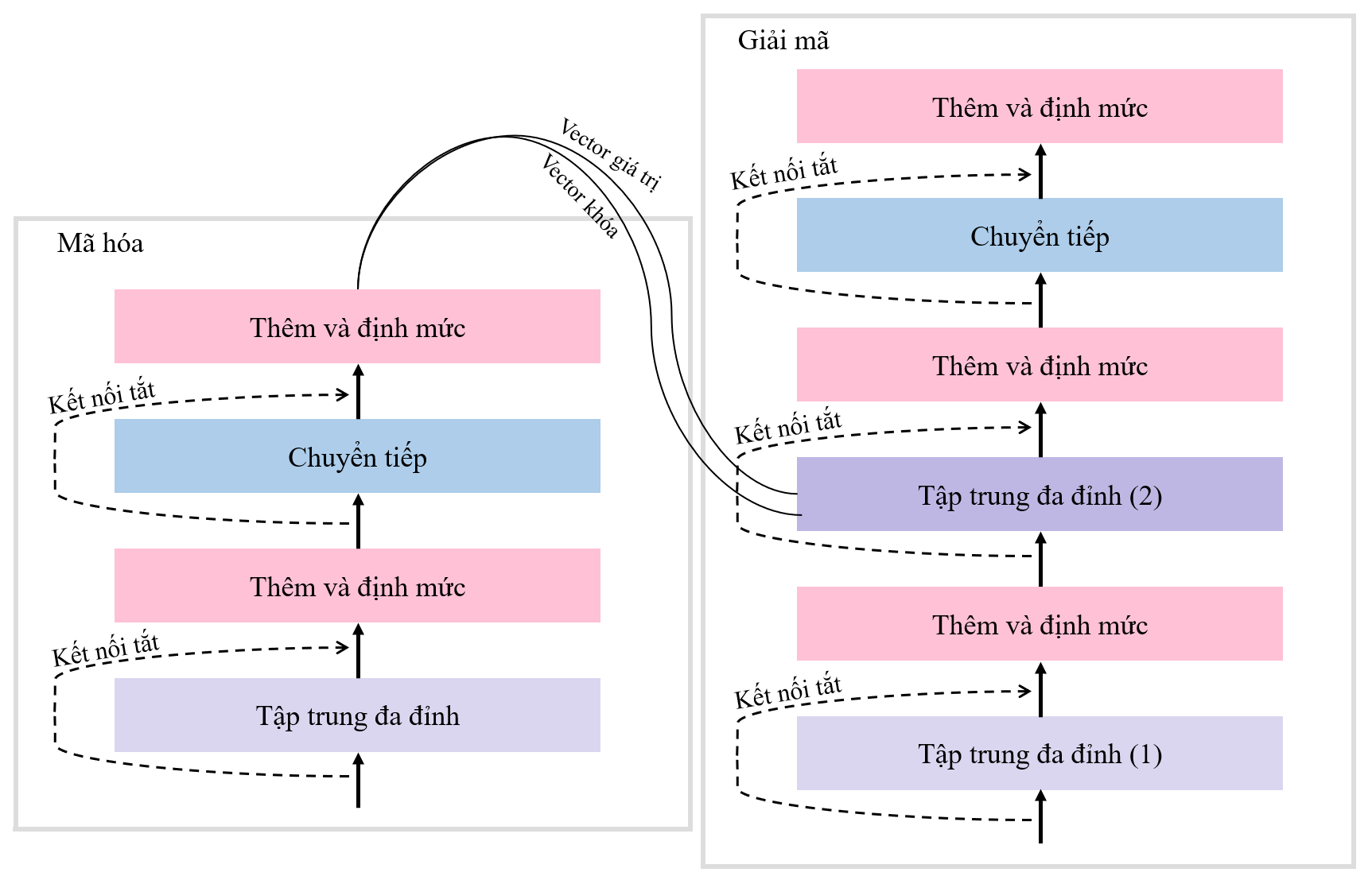
Bản đồ đặc trưng của mô hình mạng tích chập sau khi được làm phẳng thì truyền vào làm đầu vào của mô hình LSTM. Tại mỗi thời điểm, mô hình LSTM cần dự đoán từ tiếp theo trong ảnh là gì. Phần mã hóa gồm nhiều lớp mã hóa tương tự nhau. Mỗi lớp gồm 2 thành phần chính là tập trung đa đầu (multi head attention) và mạng chuyển tiếp (feedforward network). Phần mã hóa đầu tiên sẽ nhận ma trận biểu diễn các từ đã được cộng với thông tin vị trí thông qua mã hóa vị trí (positional encoding). Ma trận này được xử lý bởi tập trung đa đầu. Phần giải mã thực hiện chức năng giải mã vector của câu nguồn thành câu đích, nhận thông tin từ phần mã hóa là 2 vector khóa và giá trị. Kiến trúc của phần giải mã gần giống với phần mã hóa, nhưng có thêm một cơ chế tập trung đa đầu nằm ở giữa dùng để học mối liên hệ giữa các từ đang được dịch với các từ ở câu nguồn.

## Mô hình Transformer OCR

Kiến trúc tổng quan mô hình Transformer gồm 2 phần là phần mã hóa (encoder) và phần giải mã (decoder) [5]. Phần mã hóa (hình 3) dùng để học vector biểu của câu, mang thông tin chính xác nhất của câu đó. Phần giải mã (hình 4) thực hiện chức năng chuyển vector biểu diễn kia thành ngôn ngữ đích.



**Hình 3**. Phần mã hóa



**Hình 4**. Phần giải mã

Ưu điểm của mô hình này là khả năng xử lý song song cho các từ, trong khi mô hình LSTM lại xử lý tuần tự.

## Sử dụng mạng nơron tích chập CNN nhận dạng chữ viết tiếng Việt

Từ hai mô hình đã trình bày, chúng tôi đề xuất hai phương pháp sử dụng mạng nơrơn tích chập

* Phương pháp 1, sử dụng mạng nơron tích chập kết hợp mô hình Attention OCR.
* Phương pháp 2, sử dụng mạng nơron tích chập kết hợp mô hình Transformer OCR.

Huấn luyện mô hình AttenionOCR hay TransformerOCR hoàn toàn giống với luyện mô hình sequence-to-sequence, chúng đều sử dụng cross-entropy loss để tối ưu thay vì sử dụng CTCLoss như mô hình CRNN, tức là tại mỗi thời điểm mô hình dự đoán một từ sau đó so sánh với nhãn để tính loss và cập nhật lại trọng số của mô hình.

# III. THỰC NGHIỆM MÔ HÌNH

## Cơ sở dữ liệu chữ viết

Bộ dữ liệu gồm chữ viết tay các từ, các câu văn ngắn của nhiều người khác nhau với nét chữ khác nhau. Nội dung của các chữ viết đa dạng gồm: các danh ngôn sống, hoạt động kinh tế, tự sự văn học, chính trị và các số liệu. Kèm theo đó là nhãn text chứa nội dung văn bản tương ứng với các hình ảnh.

Để thực hiện nghiên cứu nhận dạng chữ viết tay tiếng Việt, từ bộ dữ liệu ở trên, chúng tôi tiến hành xây dựng tập TRAIN để sử dụng huấn luyện mô hình hệ thống, tập TEST để đánh giá. Để đánh giá khách quan, các tập này độc lập và không trùng lặp nhau. Tổng bộ dữ liệu có gần 10.000 ảnh, trong đó 80% thực hiện cho huấn luyện mô hình và 20% cho đánh giá kết quả.

## Phương pháp truyền thống: mô hình CRNN sử dụng CTCLoss

Chúng tôi đã thử nghiệm trên mô hình CRNN. Đầu tiên là chuẩn bị dữ liệu, thực hiện các tiền xử lý như loại bỏ nhiễu ảnh và phông nền bằng thuật toán phân cụm k-means để tạo thành ảnh trắng đen rồi nhị phân hóa ảnh bằng vào công cụ imgtxtenh [6]. Sử dụng mô hình CRNN kết hợp với cơ chế tập trung. Đối dữ liệu là ảnh sẽ dùng mô hình CNN để trích xuất đặc tính, cụ thể là VGG16. Với mô hình CRNN, kết quả của VGG được truyền trực tiếp vào mô hình LSTM. Bởi vì không có nhãn cụ thể tại mỗi thời điểm từ xuất hiện tương ứng timestep trong mô hình LSTM nên phải dùng CTCLoss để tính độ lỗi. Để ra được kết quả cuối cùng, chúng ta có thể lựa chọn câu được phát sinh bằng thuật toán best path, hoặc có thể bằng thuật toán beam search.

Kết quả như sau:

**Bảng 1.** Kết quả nhận dạng (làm tròn theo số nguyên, tỷ lệ %) chữ viết tiếng Việt sử dụng mô hình CRNN

|  |  |
| --- | --- |
| **Phương pháp** | **Kết quả thử nghiệm trên tập TEST** |
| CRNN+CTCLoss | 78,26% |

Hạn chế của mô hình CRNN sử dụng CTC Loss [10] để làm mục tiêu là số lượng đối đa có thể dự đoán bằng với chiều dài x chiều cao của các đặc trưng trích ra được từ tấm ảnh (feature maps). Do đó khi điều chỉnh kiến trúc mô hình để có thể dự đoán số kí tự phù hợp với từng bộ dataset chúng ta cần cẩn thận hơn nhiều. Để giải quyết vấn đề này, mô hình Attention OCR hoặc Transformer OCR sẽ phù hợp để chúng ta có thể dễ dàng sử dụng lại mô hình đã huấn luyện cho các loại dữ liệu khác nhau.

## Nhận dạng chữ viết sử dụng 2 mô hình Transformer OCR và Attention OCR

Để huấn luyện mô hình, chúng ta sẽ sử dụng cơ sở dữ liệu phía trên gồm khoảng 10.000 ảnh. Ngoài ra, trong mô hình đã tích hợp tập chữ cái của tiếng Việt. Ở đây chúng tôi thử nghiệm với cả hai mô hình là VGG Transfomer và VGG Sequence-to-sequence để so sánh kết quả.

Hai thử nghiệm nhận dạng chữ viết tiếng Việt đã được thực hiện ứng với hai mô hình Transformer OCR và Attention OCR (mô tả ở phần II.C). Kết quả thử nghiệm được mô tả ở Bảng 1.

**Bảng 2.** Kết quả nhận dạng chữ viết tiếng Việt (làm tròn theo số nguyên, tỷ lệ %) sử dụng mạng CNN

|  |  |
| --- | --- |
| **Phương pháp** | **Kết quả thử nghiệm trên tập TEST** |
| Transformer OCR | 87,01% |
| Attention OCR | 85,00% |

# IV. KẾT LUẬN

Bài báo này trình bày phương pháp nhận dạng chữ viết tay tiếng Việt kết hợp sử dụng mô hình mạng nơrơn tích chập (CNN) và Transformer. Nghiên cứu này bước đầu áp dụng phương pháp mạng nơron học sâu vào bài toán này. Kiến trúc mạng được chọn lựa thử nghiệm là mạng nơron tích chập CNN kết hợp mô hình Transformer OCR và kết hợp mô hình Attention OCR. Tuy nhiên, mô hình nhận dạng OCR tức việc sử dụng language model đôi lúc sẽ trở nên dư thừa. Trong tương lai, chúng tôi sẽ phát triển về khả năng phân biệt các kiểu nhận dạng, tức trường hợp cần dùng language model để phát sinh câu hoặc nhận dạng từng ký tự qua đó lựa chọn phương pháp phù hợp để giảm thiểu chi phí của công việc nhận dạng. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# TÀI LIỆU THAM KHẢO

1. Justin Johnson, Andrej Karpathy, “Convolutional Neural Networks for Visual Recognition course”, Stanford University, 2016.
2. Shifu Jain, “Bidirectional Encoder Representations for Transformers (BERT) Simplified”, 2019.
3. Arica, N.; Yarman-Vural, F.T, “Optical character recognition for cursive handwriting”, 2002.
4. Zhou, Xinyu; Yao, Cong; Wen, He; Wang, Yuzhi; Zhou, Shuchang; He, Weiran; Liang, Jiajun, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) – “EAST: An Efficient and Accurate Scene Text Detector”, 2017.
5. Jakob Uszkoreit, “Transformer: A Novel Neural Network Architecture for Language Understanding”, 2017.
6. Mauricio Villegas, “imgtxtenh - Tool for enhancing noisy scanned text images”, 2018.
7. Quốc Phạm, “VietOCR - Nhận Dạng Tiếng Việt Sử Dụng Mô Hình Transformer và AttentionOCR”, 2020.
8. Quốc Phạm, “Nhận Dạng Chữ Tiếng Việt - Vietnamese OCR”, 2018.
9. Truong Quang Vinh, Le Hoai Duy, Nguyen Thanh Nhan, “Vietnamese handwritten character recognition using convolutional neural network”, IAES International Journal of Artificial Intelligence (IJ-AI) Vol. 9, No. 2, June 2020, pp. 276-283.
10. Nguyễn Hồng Quang, Trinh Văn Loan, Phạm Ngọc Hưng, “Nhận dạng phương ngữ tiếng việt sử dụng mạng nơron tích chập CNN”, 2017.
11. Pham Ba Cuong Quoc, “Vietnamese Handwritten Characters Recognition - Convolution Recurrent Neural Nets”, 2019.

**Vietnamese handwritten character recognition using Transformer and Attention OCR Model**

# Ngoc-Tin Vo, Quang-Huy Pham

***ABSTRACT****: This paper presents the method of Vietnamese handwritten character recognition using convolutional neural network. (CNN) combined with Transformer model for language understanding. This association model has good generalizability and high accuracy on a new dataset even though it has not been trained. The test results show that the method has achieved superior results compared with the CRNN model using CTCLoss.*