BAN CƠ YẾU CHÍNH PHỦ
AN TOÀN Thông tin
INFORMATION SECURITY JOURNAL

Nhận diện văn bản tiếng Việt dựa trên mô hình Trandasformer

Thời gian gần đây, các kỹ thuật nhận diện ký tự quang học đã có bước tiến lớn với sự xuất hiện của các phương pháp mới như Transformer dành cho các ngôn ngữ Latinh. Tuy nhiên, do bản chất phức tạp của chữ viết tiếng Việt nên việc nghiên cứu về các ngôn ngữ riêng biệt này vẫn còn khá hạn chế. Điều này gây ra những thách thức đặc thù đối với OCR. OCR tiếng Việt rất quan trọng trong nhiều ứng dụng như quản lý tài liệu, lưu trữ số và nhập liệu tự động. Nghiên cứu này chỉ ra một mạng nơron sâu sử dụng kiến trúc Transformer để nhận diện từ tiếng Việt, cho ra các kết quả khả quan.

Hiệu quả của phương pháp này được đánh giá bằng cách hiệu chỉnh mô hình Transformer với kết quả chính xác đạt khoảng 95%. Kết quả này khá tốt so với các phương pháp trước đó. Điều này nhấn mạnh tiềm năng của các phương pháp dựa trên Transformer đối với OCR tiếng Việt.

TỔNG QUAN

Nhận diện ký tự quang học (Optical Character Recognition - OCR) là một công nghệ tự động nhận diện văn bản trong các hình ảnh tài liệu, sau đó chuyển nó thành văn bản để có thể tìm kiếm và chỉnh sửa được trên máy tính. Các phần mềm OCR có rất nhiều ứng dụng, bao gồm nhận diện biển số xe, đọc séc ngân hàng, xác minh chữ ký và giải mã CAPTCHA. Việc triển khai hệ thống OCR có thể gặp nhiều thách thức do sự khác biệt về phong cách viết, kích thước phông chữ, chất lượng tài liệu, bao gồm tài liệu viết tay, in hoặc quét. Những hệ thống này có thể đơn ngữ hoặc đa ngữ, hoạt động offline hoặc online. Các hệ thống OCR offline chấp nhận đầu vào ở dạng tài liệu đã được quét, in ấn hoặc viết tay, trong khi hệ thống OCR online xử lý và phân tích hình ảnh theo thời gian thực. Các ứng dụng offline bao gồm việc đọc địa chỉ bưu điện, kiểm tra séc và xử lý biểu mẫu, trong khi các bút kỹ thuật số hỗ trợ người khiếm thị hoặc người không biết chữ sử dụng hệ thống online.

Tiếng Việt là một ngôn ngữ Latinh với bảng chữ cái gồm 29 chữ cái và 5 dấu thanh, mỗi dấu có thể xuất hiện ở trên hoặc dưới các chữ cái, thay đổi hoàn toàn cách phát âm của từ. Sự phức tạp của các dấu thanh này cùng với việc một số dấu thanh và chữ cái có hình dáng tương tự, tạo ra thách thức lớn khi thiết kế hệ thống OCR cho tiếng Việt. Thêm vào đó, tiếng Việt cũng có những ký tự và dấu câu riêng biệt, làm tăng thêm mức độ khó cho các hệ thống OCR.

PHƯƠNG PHÁP TRANSFORMER ORC

Phương pháp nhận diện ký tự dựa trên mô hình Transformer là một giải pháp tiên tiến cho bài toán nhận dạng ký tự quang học. Trong đó, mô hình Transformer được sử dụng cho cả hai nhiệm vụ phân tích hình ảnh và sinh chuỗi ký tự. Phương pháp này tận dụng kiến trúc Transformer để thay thế các phương pháp truyền thống sử dụng mạng nơron tích chập (Convolutional Neural Network - CNN) và mạng nơron hồi quy (Recurrent Neural Network - RNN), nhằm xử lý đồng thời cả hình ảnh và ngôn ngữ.

Cấu trúc của mô hình gồm hai phần chính: bộ mã hóa (encoder) và bộ giải mã (decoder). Bộ mã hóa có nhiệm vụ xử lý và trích xuất đặc trưng từ hình ảnh văn bản, trong khi bộ giải mã sẽ dựa trên các đặc trưng này để tạo ra chuỗi từ tương ứng.

Bộ mã hóa (Encoder)

Bộ mã hóa bắt đầu với việc tiếp nhận hình ảnh đầu vào $x_{img} \in R^{3\times H_0 \times W_o}$ và chuyển đổi nó về kích thước cố định (H, W). Vì Transformer không thể xử lý trực tiếp hình ảnh dưới dạng lưới điểm ảnh (pixel grid), hình ảnh được chia thành các mảnh (patch) có kích thước (P, P), sao cho tổng số mảnh là $N = \frac{HW}{P^2}$. Các mảnh này sau đó được làm phẳng thành các vector và chiếu tuyến tính thành các vector có kích thước D, gọi là các patch embedding.

Công thức chiếu tuyến tính các mảnh thành vector có thể được mô tả như sau:

$$E_{patch} = Linear(Flatten(x_{img}))$$

Các vector embedding này sau đó được cộng thêm embedding vị trí để giữ lại thông tin về vị trí của mỗi mảnh trong hình ảnh gốc:

$$E_{input} = E_{patch} + E_{pos}$$

Bộ mã hóa Transformer sau đó áp dụng cơ chế tự chú ý (self-attention) và các tầng truyền tiếp để xử lý các embedding này.

Bộ giải mã (Decoder)

Bộ giải mã trong mô hình sử dụng kiến trúc Transformer chuẩn bao gồm các tầng self-attention, các tầng attention giữa bộ mã hóa và giải mã. Điểm khác biệt của bộ giải mã là nó sinh dần các token ký tự từ chuỗi đầu ra trước đó, bắt đầu từ token "[BOS]" (Beginning of Sentence).

Mỗi token đầu ra $Token_i$ tại vị trí i được tính toán dựa trên embedding của token đó và các token trước đó:

$$h_i = Proj(Emb(Token_i))$$

Sau đó, một phép biến đổi softmax được áp dụng để tính xác suất trên từ vựng:

$$\sigma(h_{ij}) = \frac{e^{h_{ij}}}{\sum_{k=1}^{V} e^{h_{ik}}}$$

với V là kích thước của từ vựng.

Bộ giải mã cũng sử dụng cơ chế chú ý giữa bộ mã hóa và giải mã để kết hợp thông tin từ đặc trưng hình ảnh được trích xuất bởi bộ mã hóa. Các đầu ra của bộ giải mã được sinh dần thông qua một quá trình dự đoán tuần tự, và quá trình này tiếp tục cho đến khi mô hình sinh ra token "[EOS]" (End of Sentence).

Huấn luyện và tiền xử lý

Phương pháp nhận diện ký tự dựa trên mô hình Transformer không yêu cầu các bước tiền xử lý hoặc hậu xử lý phức tạp. Quá trình huấn luyện diễn ra theo dạng giám sát, với chuỗi đầu ra được trích xuất từ hình ảnh văn bản và được đối chiếu với chuỗi từ đúng trong tập huấn luyện. Mô hình sử dụng hàm mất mát cross-entropy để tối ưu hóa:

$$L = -\sum_{i=1}^{T} logP(y_i|y_{1:i-1}, x_{img})$$

Trong đó, T là độ dài của chuỗi ký tự, $P(y_{1:i-1}, x_{img})$ là xác suất của token y_i tại thời điểm i, dựa trên các token trước đó và hình ảnh đầu vào.

THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

Các tham số đánh giá

Trong đánh giá mô hình OCR sử dụng kiến trúc Transformer (TransformerOCR), có hai thước đo quan trọng là Accuracy full Sequence và Accuracy per Character. Cả hai thước đo này cung cấp cái nhìn chi tiết về hiệu suất của mô hình trong việc nhận dạng dữ liệu từ hình ảnh.

Accuracy full Sequence [7] là một thước đo đánh giá độ chính xác của mô hình trong việc nhận dạng toàn bộ chuỗi ký tự (sequence) trong một hình ảnh.

Để tính Accuracy per Sequence cần sử dụng công thức sau:

$$Accuracy\ full\ Sequence = \frac{Số\ lượng\ hình\ ảnh\ được\ nhận\ dạng\ đúng}{Tổng\ số\ lượng\ hình\ ảnh\ trong\ tập\ kiểm\ tra}$$

Trong đó, số lượng hình ảnh được nhận dạng đúng là số lượng hình ảnh trong tập kiểm tra mà toàn bộ chuỗi ký tự được nhận dạng đúng. Tổng số lượng hình ảnh trong tập kiểm tra là tổng số lượng hình ảnh trong tập kiểm tra.

Accuracy per Character là một thước đo khác đánh giá độ chính xác của mô hình trong việc nhận dạng từng ký tự (character) riêng lẻ trong một hình ảnh.

Để tính Accuracy per Character, ta sử dụng công thức sau:

$$Accuracy\ per\ Character = \frac{Số\ lượng\ ký\ tự\ được\ nhận\ dạng\ đúng}{Tổng\ số\ lượng\ ký\ tự\ trong\ tập\ kiểm\ tra}$$

Tổng số lượng ký tự trong tập kiểm tra gồm: Số lượng ký tự được nhận dạng đúng là số lượng ký tự trong tập kiểm tra mà mô hình nhận dạng chính xác; Tổng số lượng ký tự trong tập kiểm tra là tổng số lượng ký tự trong tập kiểm tra.

Hai thước đo "Accuracy per Sequence" và "Accuracy per Character" cung cấp cái nhìn toàn diện về hiệu suất của mô hình OCR dựa trên kiến trúc TransformerOCR. Accuracy per Sequence đánh giá khả năng đọc và hiểu thông tin trong các đoạn văn bản dài, trong khi Accuracy per Character đánh giá khả năng phân biệt và nhận dạng chính xác từng ký tự đơn lẻ. Khi đánh giá mô hình, cần xem xét cả hai thước đo này để có cái nhìn đầy đủ và chi tiết về hiệu suất của mô hình.

Tỷ lệ lỗi ký tự (Character Error Rate - CER) là một độ đo được sử dụng trong bài toán OCR để đánh giá độ chính xác của việc nhận dạng văn bản. Nó đo lường tỉ lệ phần trăm của các ký tự bị nhận dạng sai so với số lượng tổng ký tự trong văn bản gốc. CER là một trong những phương pháp chính được sử dụng để đánh giá hiệu suất của các mô hình OCR.

CER được tính theo công thức sau:

$$CER = \frac{S + R + I}{N} \; x \; 100\%$$

Trong đó:

- *S* là số lần thay thế (số lượng ký tự bị thay đổi so với văn bản tham chiếu).
- *D* là số lần xóa (số lượng ký tự bị bỏ qua trong văn bản đầu ra so với văn bản tham chiếu).
- *I* là số lần chèn (số lượng ký tự được thêm vào văn bản đầu ra so với văn bản tham chiếu).
- *N* là tổng số ký tự trong văn bản tham chiếu.

Mẫu số *N* có thể được tính bằng công thức: $N = S + D + C$ (trong đó *C* là số ký tự được nhận diện đúng).

CER đánh giá độ chính xác của việc nhận dạng ký tự trong văn bản bằng cách tính tỉ lệ phần trăm của các ký tự bị nhận dạng sai so với tổng số ký tự trong văn bản gốc. Nó cho biết mức độ chính xác của mô hình OCR và có giá trị từ 0 đến 100%. Một CER càng thấp thể hiện mô hình nhận dạng ký tự càng chính xác.

Dữ liệu sử dụng

Dữ liệu sử dụng trong quá trình đào tạo mô hình TransformerOCR bao gồm một tập hợp đa dạng các hình ảnh chứa chữ in từ nhiều tài liệu văn bản. Dữ liệu được thu thập từ nhiều tài liệu khác nhau để đảm bảo tính đa dạng và phù hợp với thực tế. Tập dữ liệu đã được gán nhãn với văn bản tương ứng trong từng hình ảnh.

Kích thước của tập dữ liệu là lớn và đủ lớn để huấn luyện mô hình TransformerOCR một cách hiệu quả. Dữ liệu được chia thành ba phần chính: tập huấn luyện, tập kiểm tra và tập xác thực. Tập huấn luyện được sử dụng để huấn luyện mô hình, tập kiểm tra dùng để đánh giá hiệu suất của mô hình trong quá trình huấn luyện và tập xác

thực dùng để điều chỉnh các siêu tham số quan trọng của mô hình. Kích thước chi tiết của từng phần được đề cập trong bảng 1.

BẢNG 1: KÍCH THƯỚC TẬP DỮ LIỆU THỰC NGHIỆM
MÔ HÌNH TRANSFORMEROCR

Tập dữ liệu	**Số lượng**
Tập huấn luyện	670
Tập kiểm tra	511
Tập xác thực	450

Mô hình ban đầu là mô hình TransformerOCR được cung cấp trong thư viện VietOCR. Mô hình này được pre-trained từ tập dữ liệu gồm 10 triệu ảnh chữ in tiếng Việt được sinh tự động. Mục đích là sử dụng dữ liệu về chữ in trong các tài liệu văn bản, tiến hành fine-tune mô hình gốc để đạt được mô hình có độ chính xác cao hơn.

Cài đặt thực nghiệm và lựa chọn siêu tham số

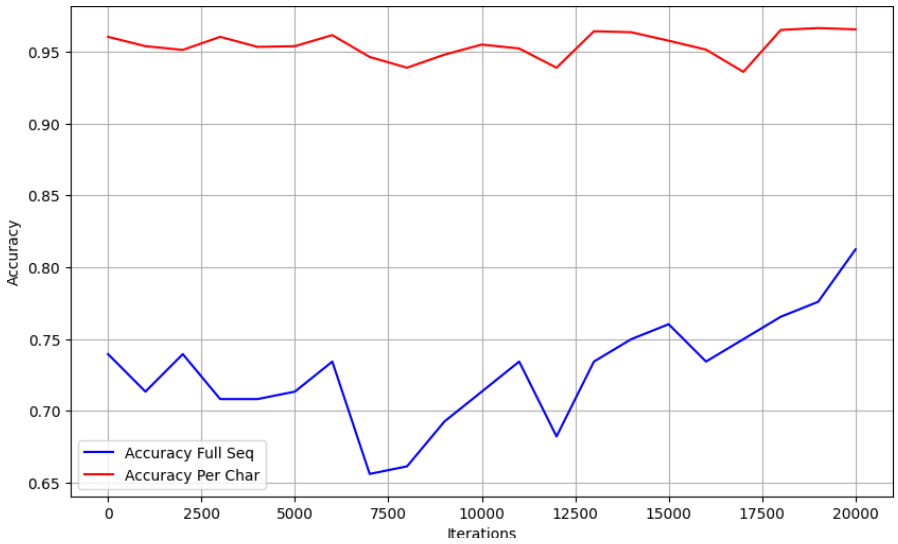
Sử dụng thư viện VietOCR trên Python để cài đặt đặt mô hình TranformerOCR. Thư viện VietOCR hỗ trợ cài đặt huấn luyện các mô hình OCR sử dụng kiến trúc mạng học sâu như Transformer.

Về siêu tham số (Hyperparameters) có:

- Số lượng đầu ra (head): 8
- Số lượng lớp mã hóa (encoder layers): 6
- Số lượng lớp giải mã (decoder layers): 6
- Số chiều của lớp Fully-Connected: 2048
- Tỷ lệ dropout cho lớp Positional Encoding: 0,1
- Tỷ lệ dropout cho lớp Attention: 0,1
- Tốc độ học tối đa (max learning rate): 0,0003
- Tỷ lệ epochs sẽ được sử dụng cho giai đoạn tăng tốc tốc độ học (learning rate warm-up): 0,1
- Kích thước của mỗi batch (batch size): 32
- Tổng số lượng bước (iterations): 20000

Kết quả thực nghiệm của mô hình TransformerOCR đã cho thấy rằng trong quá trình huấn luyện và kiểm chứng, mô hình đều hội tụ và độ chính xác của mô hình tăng so

với mô hình TransformerOCR. Theo Hình 1, ở bước cuối cùng, giá trị accuracy full sequence của mô hình đạt khoảng 81,25% và giá trị accuracy per character đạt 96,56% trong quá trình huấn luyện. Điều này làm rõ rằng mô hình đã được huấn luyện hiệu quả và phù hợp với bộ dữ liệu. Để làm rõ sự phù hợp này, mô hình sau khi huấn luyện sẽ được đánh giá và so sánh thông qua độ đo CER đã được nêu ở mục trước.



Hình 1. Giá trị độ chính xác của mô hình TransformerOCR trong quá trình huấn luyện

BẢNG 2. KẾT QUẢ ĐÁNH GIÁ TRÊN BỘ DỮ LIỆU CHỮ IN TRONG TÀI LIỆU VĂN BẢN

Mô hình	**Kiến trúc**	**CER**
TransformerOCR Fine-tuned	VGG19 - Transformer	3,44
TransformerOCR Pre-trained	VGG19 - Transformer	12,01
AttentionOCR	VGG19 - Seq2Seq	12,99

Từ Bảng 2 có thể đánh giá các mô hình OCR trên bộ dữ liệu chữ in trong các tài liệu văn bản như sau:

Thứ nhất, mô hình TransformerOCR Fine-tuned đã đạt kết quả tốt nhất với CER chỉ là 3,44. Điều này chứng tỏ mô hình đã được cải thiện hiệu quả sau quá trình fine-tuning, giúp nâng cao độ chính xác trong việc nhận dạng chữ in.

Thứ hai, mô hình TransformerOCR Pre-trained mặc dù đã được huấn luyện trước đó, nhưng vẫn có CER là 12,01, cao hơn rõ rệt so với mô hình fine-tuned. Điều này có thể chỉ ra rằng mô hình pre-trained cần phải được điều chỉnh và tinh chỉnh thêm để đạt hiệu suất tốt hơn.

Thứ ba, mô hình AttentionOCR có kết quả tệ nhất trong bảng với CER là 12,99. Điều này cho thấy mô hình này chưa phát huy hiệu quả trong việc nhận dạng chữ in so với hai mô hình TransformerOCR.

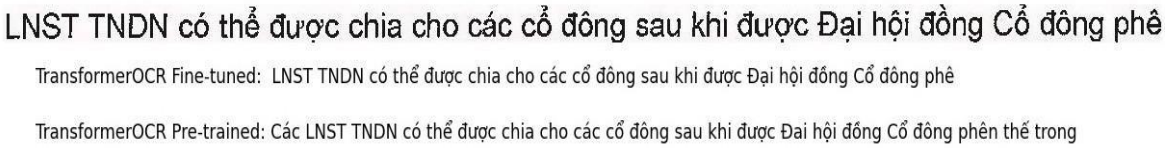
Từ bảng đánh giá có thể thấy rõ rằng việc fine-tuning mô hình TransformerOCR giúp cải thiện hiệu suất đáng kể. Ngoài ra, kiến trúc Transformer trong mô hình cũng có hiệu quả hơn so với kiến trúc Seq2Seq trong mô hình AttentionOCR. Điều này có thể giúp các ứng dụng OCR trong việc nhận dạng hình ảnh văn bản chữ in đạt được độ chính xác cao hơn và đáng tin cậy hơn.

Hình ảnh kết quả

Để trực quan hóa kết quả thử nghiệm, mô hình sẽ được kiểm chứng với hình ảnh trong tập dữ liệu thử nghiệm và so sánh với mô hình TransformerOCR pre-trained. Dữ liệu đầu vào và đầu ra được mô tả lần lượt trong Hình 2, Hình 3 và Hình 4.

Hình 2: Kết quả OCR văn bản của mô hình TransformerOCR

Hình 3: Kết quả OCR hình ảnh chứa từ của mô hình TransformerOCR



Hình 4. Kết quả OCR số của mô hình TransformerOCR

Dựa trên hình ảnh kết quả dự đoán của mô hình TransformerOCR được fine-tuned và mô hình TransformerOCR pre-trained, có thể thấy trong Hình 2 và 3 cho kết quả tốt hơn khi

thực hiện OCR, các số liệu cũng như từ thường gặp trong báo cáo tài chính, mô hình này cũng cải thiện được việc thêm các tiền tố và hậu tố vào kết quả của mô hình gốc như trong Hình 4. Chất lượng của mô hình được fine-tune có độ chính xác cao hơn.

Kết quả cho thấy họ mô hình TransformerOCR được fine-tune phù hợp với bài toán OCR trên dữ liệu hình ảnh tài liệu văn bản tiếng Việt.

KẾT LUẬN

Nhận dạng ký tự quang học là quá trình chuyển đổi văn bản từ hình ảnh thành định dạng có thể đọc được bằng máy. Trong những năm gần đây, các kỹ thuật OCR đã đạt được những tiến bộ đáng kể nhờ các phương pháp mới như Transformer. Đối với bài toán OCR, việc sử dụng mô hình Transformer đạt độ chính xác cao trong việc nhận dạng chữ in tiếng Việt từ hình ảnh, với độ chính xác lên đến 95%. Điều này góp phần cải thiện quá trình trích xuất thông tin, tự động hóa và nâng cao tính chính xác trong quá trình xử lý dữ liệu. Chỉ số CER của mô hình TransformerOCR sau khi fine-tuned đã giảm đáng kể từ 0,0087 xuống còn 0,0035, chứng tỏ hiệu quả của việc tinh chỉnh mô hình để phù hợp hơn với dữ liệu thực tế. Những kết quả này nhấn mạnh tiềm năng to lớn của các mô hình dựa trên Transformer trong việc cải thiện độ chính xác của OCR tiếng Việt, mở ra những hướng nghiên cứu và phát triển mới trong tương lai.

TÀI LIỆU THAM KHẢO

[1]. Sargur N. Srihari, Ajay Shekhawat, and Stephen W. Lam. “Optical character recognition (OCR)”. Encyclopedia of Computer Science. John Wiley and Sons Ltd., GBR, 1326–1333. (2003)

[2]. Vaswani, A. “Attention is all you need.” Advances in Neural Information Processing Systems (2017).

[3]. Li, Minghao, et al. “Trocr: Transformer-based optical character recognition with pre-trained models.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. No. 11. 2023.

[4]. Dirk Alvermann, “Word Error Rate & Character Error Rate – How to evaluate a model”, 10/2019

[5]. Q. Pham, “Vietocr – Vietnameses recognition using Transformer Model and AttentionOCR”. url: https://pbcquoc.github.io/vietocr/.

[6]. Singh, Amarjot & Bacchuwar, Ketan & Bhasin, Akshay. “A Survey of OCR Applications”. International Journal of Machine Learning and Computing (IJMLC). 10.7763/IJMLC.2012.V2.137. (2012)

[7]. Python. url: https://www.python.org.

[8]. O'Shea, K. “An introduction to convolutional neural networks.” arXiv preprint arXiv:1511.08458 (2015).

[9]. Sherstinsky, Alex. “Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network.” Physica D: Nonlinear Phenomena 404 (2020): 132306.

[10]. Zhang, Jinjin, et al. “A feasible framework for arbitrary-shaped scene text recognition.” arXiv preprint arXiv:1912.04561 (2019).