Nhận dạng văn bản tiếng Việt trong ảnh ngoại cảnh bằng học sâu

Huỳnh Văn Huy[1], Nguyễn Thị Thanh Tân[2*], Ngô Quốc Tạo[3]

[1] Trường Đại học Lạc Hồng;
[2] Trường Đại học Điện lực;
[3] Viện Công nghệ Thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.
[*]Email: tanntt@epu.edu.vn
Nhận bài: 10/7/2023; Hoàn thiện: 20/9/2023; Chấp nhận đăng: 10/10/2023; Xuất bản: 25/10/2023.
DOI: https://doi.org/10.54939/1859-1043.j.mst.90.2023.140-149

TÓM TẮT

Bài báo này đề xuất một phương pháp hiệu quả để nhận dạng văn bản tiếng Việt trong ảnh ngoại cảnh. Phương pháp được đề xuất dựa trên ý tưởng kết hợp ba luồng xử lý đồng thời trong một công đoạn nhận dạng, bao gồm: (i) Nhận dạng (dự đoán) chuỗi ký tự từ hình ảnh; (ii) Xử lý ngữ cảnh; (iii) Hợp nhất và hiệu chỉnh lỗi. Hiệu quả của phương pháp được kiểm nghiệm trên hai tập dữ liệu ảnh ngoại cảnh được thu thập từ thực tế VinText và VnSceneText. Các kết quả thực nghiệm cho thấy phương pháp có khả năng phát hiện được các văn bản có hình dạng và kích thước bất kỳ với độ chính xác cao và ổn định. Cụ thể, phương pháp đạt độ chính xác mức từ (word accuracy), độ chính xác mức ký tự (character accuracy) là (81.87%, 93.02%) và (82.56%, 94.33%) lần lượt trên hai tập dữ liệu thử nghiệm.

Từ khoá: Phát hiện; Nhận dạng; Đặc trưng; Xác suất; Độ chính xác.

1. MỞ ĐẦU

Phát hiện và nhận dạng văn bản trong ảnh ngoại cảnh (gọi một cách ngắn gọn là nhận dạng văn bản ngoại cảnh) hiện được ứng dụng rất phổ biến trong thực tế, điển hình như: Đọc bảng hiệu, hộp sản phẩm, nhãn hàng; Nhận dạng nhãn và thông số kỹ thuật trên các linh kiện điện tử - ứng dụng trong các dây chuyền sản xuất công nghiệp; Nhận dạng và dịch tên đường phố, các biển báo giao thông, biển chỉ dẫn, menu nhà hàng cho người khiếm thị hoặc du khách nước ngoài đến Việt Nam du lịch; Nhận dạng phụ đề và các số liệu từ video,...

Các phương pháp nhận dạng văn bản ngoại cảnh hiện nay được đề xuất chủ yếu dựa trên nền tảng học sâu. một số phương pháp nhận dạng văn bản dựa trên chuỗi [1, 4-7] đã sử dụng mô hình CTC [8] để dự đoán các chuỗi ký tự. Trong [9], một cửa sổ trượt trước tiên được áp dụng vào ảnh dòng văn bản (text-line image) để thu thập thông tin ngữ cảnh một cách hiệu quả, sau đó sử dụng bộ dự đoán CTC để dự đoán các từ đầu ra. Rosetta [4] chỉ sử dụng các đặc trưng được trích xuất từ mạng nơ-ron tích chập bằng cách áp dụng một mô hình ResNet [10] làm mạng backbone để dự đoán các chuỗi đặc trưng. Trong [11, 12], một cửa sổ trượt trước tiên được áp dụng vào ảnh dòng văn bản (text-line image) để thu thập thông tin ngữ cảnh một cách hiệu quả, sau đó sử dụng bộ dự đoán CTC để dự đoán các từ đầu ra. Để trích xuất thông tin ngữ cảnh tốt hơn, một số công trình [1, 4, 6] đã sử dụng RNN [13] kết hợp với CTC để xác định xác suất có điều kiện giữa chuỗi được dự đoán và chuỗi mục tiêu. Gần đây, với quan điểm rằng năng lực hạn chế của các mô hình ngôn ngữ xuất phát từ việc mô hình hóa ngôn ngữ ngầm định, biểu diễn các đặc trưng một chiều và mô hình ngôn ngữ với đầu vào nhiễu, Fang và cộng sự [14] đã đề xuất mô hình ABINet tự điều khiển (Autonomous), hai chiều (Bidirectional) và lặp lại (Iterative) để nhận dạng văn bản tiếng Anh và tiếng Trung trong ảnh ngoại cảnh. Các kết quả thực nghiệm cho thấy ABINet có khả năng giải quyết tốt các trường hợp ảnh đầu vào có chất lượng thấp và đạt độ chính xác cao nhất so với các phương pháp nhận dạng văn bản ngoại cảnh hiện có trên các tập dữ liệu tiếng Anh đã được công bố.

Đối với việc nhận dạng văn bản ngoại cảnh tiếng Việt, ngoài việc phải đối mặt với các thách thức của bài toán nhận dạng văn bản ngoại cảnh nói chung như đã đề xuất ở trên, còn gặp phải

các khó khăn về tầng dấu mũ và dấu thanh điệu trong tiếng Việt. Các tầng dấu này có thể gây ra vấn đề dính chữ và dính dòng trong văn bản, làm giảm độ chính xác trong phát hiện và nhận dạng. Do đó, để đảm bảo độ chính xác nhận dạng cần có thêm tích hợp các cơ chế xử lý riêng biệt phù hợp với các đặc thù của tiếng Việt. Khảo sát thực tế cho thấy các kết quả nghiên cứu trong phát hiện và truy xuất văn bản tiếng Việt trong ảnh ngoại cảnh hiện còn rất hạn chế. Một vài kết quả nghiên cứu tiêu biểu trong nhận dạng văn bản ngoại cảnh tiếng Việt đã được công bố gần đây [14, 16].

Bài báo này đề xuất một phương pháp hiệu quả để nhận dạng văn bản tiếng Việt trong ảnh ngoại cảnh (gọi một cách ngắn gọn là nhận dạng văn bản ngoại cảnh tiếng Việt). Mô hình nhận dạng của phương pháp đề xuất được chia thành 03 khối chức năng chính: Khối nhận dạng/dự đoán chuỗi ký tự, khối xử lý ngữ cảnh, khối hợp nhất và hiệu chỉnh lỗi. Cấu trúc của khối thứ nhất bao gồm một mạng backbone Resnet-50, một mạng mô hình hóa chuỗi, một module tập trung vị trí (position attention) và một bộ phân lớp (C1) để dự đoán các ký tự trên chuỗi đầu vào cần nhận dạng. Khối thứ hai có vai trò xử lý thông tin ngữ cảnh, khối này bao gồm mô hình ngôn ngữ tiếng Việt hai chiều (bidirectional) mức ký tự và mô hình dự đoán xác suất. Khối thứ 3 bao gồm một mô hình hợp nhất và một bộ dự đoán ký tự (C2). Mô hình hợp nhất có vai trò kết hợp các thông tin ngữ cảnh (đặc trưng ngôn ngữ, xác suất dự báo của các ký tự) với tập ánh xạ đặc trưng và từ điển ứng cử viên để dự đoán ký tự.

Các đóng góp chính của chúng tôi trong bài báo này bao gồm:

Thứ nhất, chúng tôi đã đề xuất một phương pháp hiệu quả để nhận dạng văn bản ngoại cảnh tiếng Việt đạt độ chính xác cao, không bị ảnh hưởng nhiều bởi nhiễu cũng như hướng của văn bản. Phương pháp có khả năng thích nghi và giải quyết tốt đối với các văn bản có hình dạng bất kỳ bao gồm cả các văn bản cong (curve text). Thứ hai, chúng tôi đã thiết kế các lớp mạng tương ứng với các mô hình ngôn ngữ mức ký tự (character level) và mức từ (word level) để biểu diễn thông tin ngữ cảnh, sử dụng trong quá trình nhận dạng. Thứ ba, chúng tôi đã thu thập và xây dựng được một tập dữ liệu gồm 3000 ảnh ngoại cảnh tiếng Việt từ môi trường thực tế, phục vụ cho quá trình nghiên cứu và đánh giá thử nghiệm các thuật toán phát hiện và nhận dạng văn bản tiếng Việt trong ảnh ngoại cảnh.

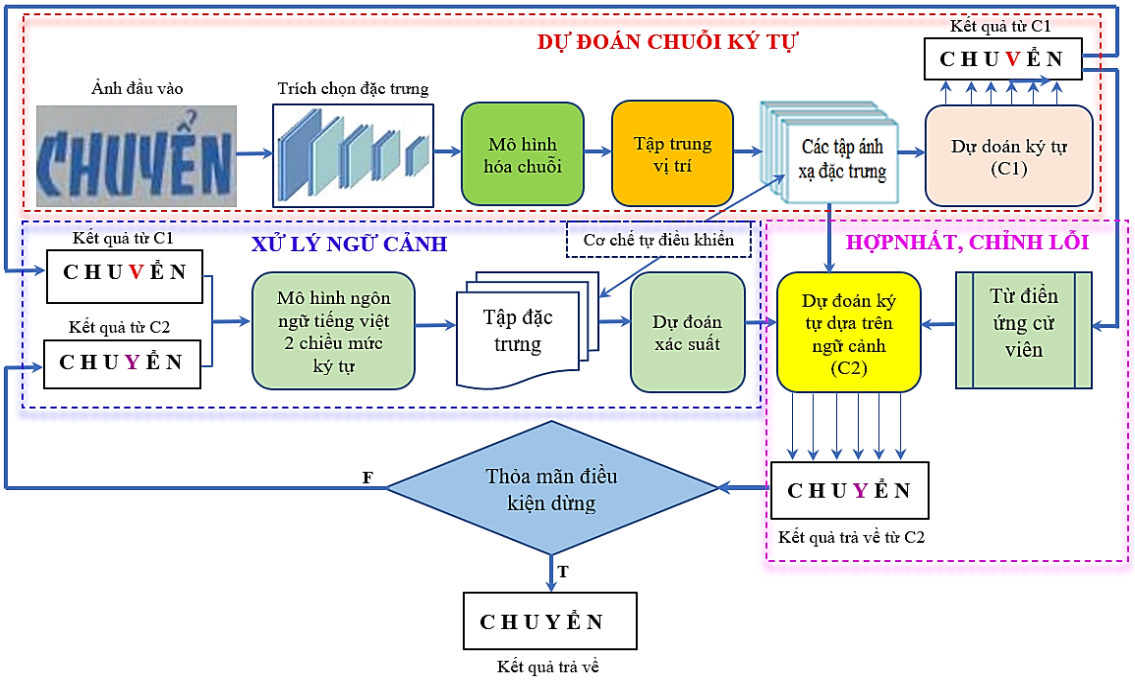
Các nội dung còn lại của bài báo được trình bày như sau: Phần 2 trình bày tóm lược các hướng tiếp cận liên quan. Ý tưởng chính và các bước thực hiện của phương pháp đề xuất được mô tả chi tiết trong phần 3. Phần 4 trình bày quá trình kiểm thử và đánh giá hiệu quả của phương pháp. Một số kết luận và hướng phát triển được đề cập trong phần 5.

2. PHƯƠNG PHÁP

Kế thừa các ý tưởng chính từ ABINet [14], chúng tôi đề xuất phương pháp VNSTR (Vietnamese Scene Text Regnition) để nhận dạng văn bản (bước thứ 3 trong quy trình trên). Đầu vào của mô hình là hình ảnh của một từ hoặc một chuỗi ký tự, đầu ra là từ hoặc chuỗi ký tự tương ứng được nhận dạng. Phương pháp này tập trung vào hai mục tiêu chính: i) Có khả năng xử lý tốt đối với các tầng dấu mũ và dấu thanh điệu trong tiếng Việt; ii) Đạt độ chính xác nhận dạng cao và ổn định trên các loại ảnh văn bản ngoại cảnh đầu vào có chất lượng, hình dạng và chủng loại font chữ bất kỳ. Các thực hiện cơ bản của phương pháp được mô tả cụ thể trên hình 1.

Quy trình thực hiện của phương pháp đề xuất như sau: Mỗi ảnh đầu vào cần nhận dạng trước tiên sẽ được đưa qua một mạng backbone để trích chọn đặc trưng. Kết quả đầu ra của mạng này sẽ tiếp tục được đưa qua mạng tập trung vị trí (position attention) để lấy được các thông tin ngữ cảnh quan trọng của các đặc trưng. Tập ánh xạ đặc trưng thu được từ bước này sẽ được sử dụng làm đầu vào đồng thời cho cả bộ phân lớp C1 và C2. Hàm softmax trong bộ phân lớp C1sẽ tiến hành phân lớp dựa trên tập ánh xạ đặc trưng đầu vào, trả về kết quả là chuỗi ký tự tương ứng nhận dạng được của ảnh đầu vào. Dựa trên kết quả trả về của bộ phân lớp C1, một danh sách ứng

cử viên sẽ được sinh ra và tự động cập nhật vào từ điển ứng cử viên. Bộ phân lớp C2 sẽ tiến hành hợp nhất các ánh xạ đặc trưng đầu vào với các thông tin ngữ cảnh thu được từ mô hình dự đoán xác suất và tham chiếu tới từ điển ứng cử viên để ra quyết quyết định.



Hình 1. *Phương pháp nhận dạng từ tiếng Việt trong ảnh ngoại cảnh.*

Từ điển ứng cử viên ở đây được sử dụng với mục đích hạn chế các nhiễu đầu vào. Kết quả trả về từ bộ phân lớp C1 và C2 tiếp tục được sử dụng làm đầu vào của mạng BCN để trích xuất đặc trưng ngôn ngữ và dự đoán xác xuất của các ký tự. Sau đó, quá trình hiệu chỉnh và phân lớp tiếp tục lặp lại cho tới khi kết quả nhận dạng thỏa mãn.

2.1. Dự đoán chuỗi ký tự từ hình ảnh

Mô hình dự đoán chuỗi ký tự từ hình ảnh bao gồm bốn công đoạn: Trích chọn đặc trưng; Mô hình hóa chuỗi; Tăng cường thông tin ngữ cảnh; và dự đoán/nhận dạng ký tự. Mô hình này chịu trách nhiệm phân tích và xử lý hình ảnh văn bản, sau đó chuyển đổi thành một chuỗi các ký tự tiềm năng. Điểm mạnh của mô hình này nằm ở khả năng xử lý đầu vào dạng hình ảnh và giữ sự linh hoạt trong việc nhận diện các ký tự qua nhiều loại font chữ khác nhau, kích thước và định dạng. Để trích chọn đặc trưng từ ảnh đầu vào, chúng tôi sử dụng kiến trúc mạng backbone ResNet-50 [10] làm backbone. Đây là một kiến trúc mạng backbone được đánh giá là cấu trúc mạng backbone tốt nhất và được sử dụng phổ biến nhất hiện nay. Kiến trúc này gồm 50 lớp (layer): Zero Padding (Lớp đệm), CONV (Lớp tích chập - Convolution), Max poolling, các Conv Block (Khối tích chập), các ID Block (Khối định danh, Avg Pooling (Average Pooling) và Flattening. Mỗi lớp tích chập trong Resnet đều được áp dụng kỹ thuật chuẩn hóa Batch Normalization (BatchNorm). Quá trình thực hiện của mạng được chia thành 6 công đoạn (stage), được ký hiệu từ stage 1 đến stage 6. Kết quả thực hiện của mỗi stage là các ánh xạ đặc trưng chuỗi (Feature maps):

$$\mathbf{F}_b = \mathcal{B}ackbone(\boldsymbol{x}) \in \mathbb{R}^{h \times w \times d} \quad (1)$$

Trong đó: h, w là kích thước (chiều cao, chiều rộng) của ánh xạ đặc trưng, được xác định tương ứng bằng ¼ kích thước (chiều cao, chiều rộng) của ảnh đầu vào; d là số chiều của đặc trưng.

Để mô hình hóa chuỗi ký tự, chúng tôi sử dụng kiến trúc Transformer được đề xuất trong

[17]. Kiến trúc này có khả năng giải quyết tốt các vấn đề cần xác định mối quan hệ giữa các phần tử cách xa nhau trong chuỗi, mà các kiến trúc sâu truyền thống như RNN hoặc LSTM gặp khó khăn. Một kiến trúc Transform gồm hai phần chính: encoder (mã hóa) và decoder (giải mã). Encoder nhận đầu vào là các ánh xạ đặc trưng chuỗi $\mathbf{F}_b$ thu được từ mạng backbone và chuyển đổi nó thành một chuỗi biểu diễn ẩn. Encoder gồm nhiều lớp, mỗi lớp lại chứa hai sub-layer là multi-head self-attention và position-wise feed-forward. Decoder sau đó nhận chuỗi biểu diễn ẩn từ encoder và dự đoán chuỗi đầu ra. Decoder cũng có cấu trúc tương tự encoder nhưng thêm một sub-layer attention cho việc định hướng đầu ra (output). Cả hai phần này sử dụng cơ chế tập trung (attention), cho phép mô hình tập trung vào các phần khác nhau của chuỗi đầu vào khi dự đoán đầu ra. Điều này giúp kiến trúc Transformer vượt trội hơn so với kiến trúc truyền thống trong việc xử lý các chuỗi dài và biến đổi ngữ cảnh.

$$\mathbf{F}_m = Transformer(\mathbf{F}_b) \in \mathbb{R}^{h \times w \times d} \quad (2)$$

Module tập trung vị trí (position attention) được sử dụng để xác định mức độ ảnh hưởng (hoặc mức độ tập trung) của một phần tử tới các phần tử khác trong cùng chuỗi. Module này nhận đầu vào là các ánh xạ chuỗi F_m và trả về đầu ra là các ánh xạ đặc trưng đã được tăng cường thông tin ngữ cảnh theo vị trí. Thông qua việc thêm thông tin về vị trí vào dữ liệu, mô hình có thể hiểu được ý nghĩa bối cảnh và trình tự của các phần tử trong một chuỗi. Ở đây, từ mô hình Transformer, thông tin vị trí được mã hóa thành các véc-tơ vị trí và được cộng vào các véc-tơ đầu vào. Khi đó, cơ chế position attention cho phép tập trung vào các phần tử quan trọng liên quan đến phần tử đang xét, dựa trên vị trí và ý nghĩa của chúng trong chuỗi.

Module dự đoán ký tự tiến hành chuyển đổi các đặc trưng hình ảnh thành xác suất ký tự một cách song song, dựa trên mô hình truy vấn:

$$\mathbf{F}_a = \text{softmax}\left(\frac{\mathbf{PE} \times \mathbf{ED}^{\top}}{\sqrt{d}}\right) \times \mathbf{IM} \quad (3)$$

Trong đó: PE, ED và IM là các ánh xạ đặc trưng (biểu diễn dưới dạng tensor), được xác định như sau:

$$\mathbf{PE} = Position\_Encoder(t) \in \mathbb{R}^{t \times d} \quad (4)$$

$$\mathbf{ED} = Encoder\_Decoder(\mathbf{F}_m) \in \mathbb{R}^{h \times w \times d} \quad (5)$$

$$\mathbf{IM} = Identity_{Mapping(\mathbf{F}_m)} \in \mathbb{R}^{h \times w \times d} \quad (6)$$

PE là ánh xạ mã hóa vị trí (positional encoding) của thứ tự phần tử (ký tự), t là độ dài của chuỗi ký tự; *ED* là kết quả thực hiện của mạng $Encoder\_Decoder$. Mạng này có kiến trúc tương tự như mạng Unet [18]. Các tham số của mạng này được mô tả cụ thể trên bảng 1; *IM* là một ánh xạ đồng nhất (Identity Mapping).

2.2. Xử lý thông tin ngữ cảnh

2.2.1. Cơ chế tự điều khiển

Chiến lược tự điều kiển bao gồm các đặc điểm sau đây: 1) Mô hình ngôn ngữ được coi là một mô hình độc lập với các thao tác chỉnh sửa chính tả, nhận các vector xác suất của ký tự làm đầu vào và đưa ra các phân phối xác suất của các ký tự dự kiến; 2) Luồng gradient huấn luyện bị chặn (BGF) tại các vector đầu vào; 3) Mô hình ngôn ngữ được huấn luyện riêng biệt từ dữ liệu văn bản không được gán nhãn (unlabeled text data).

Cho một chuỗi văn bản $\boldsymbol{s}_{1:n} = \{s_1, \ldots, s_n\}$ với độ dài n, các phương pháp dựa trên cơ chế tập trung về cơ bản thực hiện mô hình hóa ngôn ngữ một cách ngầm định, tức là $P(s_i \mid \boldsymbol{s}_{1:i-1}, \mathcal{F}_v)$, trong đó $\mathcal{F}_v$ là các đặc trưng hình ảnh. $\mathcal{F}_v$ có thể là một yếu tố sai lệch từ khía cạnh của mô hình ngôn ngữ, vì $P(y_i)$ được ước lượng một phần dựa trên $\mathcal{F}_v$. Để khắc phục điều đó, chiến lược tự điều khiển định nghĩa mô hình ngôn ngữ với giá trị xác suất $P(s_i \mid \boldsymbol{s}_{1:i-1})$ được tính một cách chính xác bằng cách tách riêng mô hình dự đoán chuỗi ký tự với mô hình ngôn ngữ. Mặt khác,

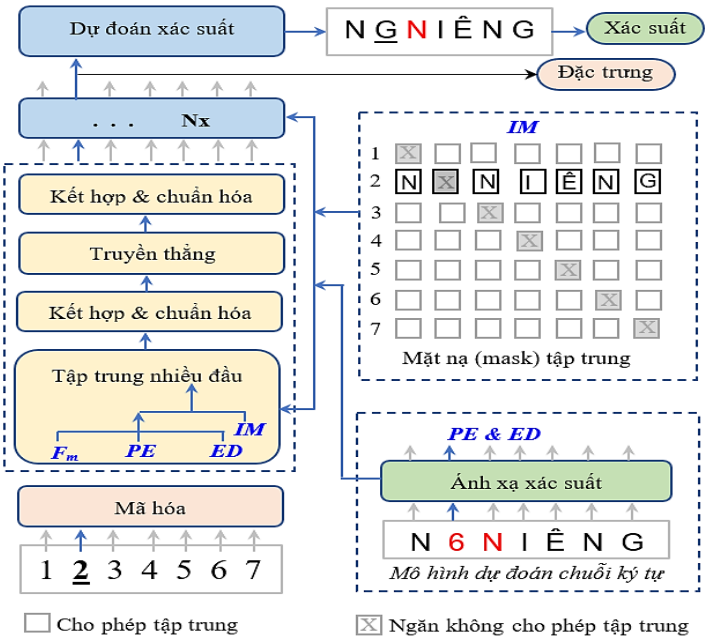
do các ký tự ngữ cảnh $\boldsymbol{s}_{1:i-1}$được tính trực tiếp từ $\mathcal{F}_v$, yếu tố sai lệch vẫn tồn tại trong quá trình lan truyền ngược. Do đó, cơ chế BGF (Blocked Gradient Flow) được sử dụng để đảm bảo tính độc lập của việc huấn luyện mô hình dự đoán chuỗi và mô hình ngôn ngữ.

2.2.2. Mô hình ngôn ngữ tiếng Việt hai chiều mức ký tự

Cho một chuỗi văn bản $\boldsymbol{s} = (s_1, \dots, s_n)$ với độ dài văn bản n và số lớp c, xác suất điều kiện của s_i cho các mô hình ngôn ngữ mức ký tự hai chiều và một chiều là:

$$P(s_i \mid s_n, \dots, s_{i+1}, s_{i-1}, \dots, s_1)$$

và $\quad P(s_i \mid s_{i-1}, \dots, s_1)$



Hình 2. *Biểu diễn mô hình ngôn ngữ tiếng Việt hai chiều mức ký tự.*

Từ góc độ lý thuyết thông tin, entropy (độ hỗn độn thông tin) của một biểu diễn hai chiều có thể được định lượng là $H_y = (n-1) \times \log\, c$. Tuy nhiên, đối với một biểu diễn một chiều thì thông tin là $\frac{1}{n}\sum_{i=1}^{n}\ (i-1)\log\, c = \frac{1}{2}H_y$. Chúng ta thấy rằng, các phương pháp trước đây thường sử dụng một mô hình kết hợp (ensemble model) từ hai mô hình một chiều. Biểu diễn một chiều này về cơ bản thu được $\frac{1}{2}H_y$ thông tin, dẫn đến khả năng trừu tượng hóa đặc trưng bị giới hạn so với biểu diễn hai chiều. Để khắc phục điều đó, chúng tôi sử dụng kiến trúc mạng BCN (bidirectional cloze network) trong [14] để biểu diễn mô hình ngôn ngữ hai chiều mức ký tự tiếng Việt (hình 2).

Khác với Transformer thông thường, các vectơ ký tự được đưa vào các khối tập trung đa đầu vào thay vì tầng đầu tiên của mạng. Ngoài ra, attention mask trong multi-head attention được thiết kế để tránh "nhìn thấy chính nó". Bên cạnh đó, trong BCN không có cơ chế self-attention để tránh rò rỉ thông tin qua các bước thời gian. Cơ chế tập trung bên trong các khối multi-head attention có thể hình thức hóa như sau:

$$\mathbf{M}_{ij} = \begin{cases} 0, & i \neq j \\ -\infty, & i = j' \end{cases} \tag{7}$$

$$\mathbf{K}_i = \mathbf{V}_i = P(s_i)\mathbf{W}_l \tag{8}$$

$$\mathbf{F}_{mha} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}} + \mathbf{M}\right)\mathbf{V} \tag{9}$$

Trong đó: $\mathbf{Q} \in \mathbb{R}^{t\times d}$ là mã hóa vị trí của ký tự trong lớp đầu tiên và đầu ra của lớp cuối cùng. $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{t\times d}$ được xác định từ xác suất ký tự $P(s_i) \in \mathbb{R}^{c}$; $\mathbf{W}_l \in \mathbb{R}^{c\times d}$ là ma trận ánh xạ có thể huấn luyện được. $\mathbf{M} \in \mathbb{R}^{t\times t}$ là attention mask, được sử dụng để ngăn không cho tập trung vào ký tự hiện tại. Sau khi xếp chồng các tầng BCN thành kiến trúc sâu, biểu diễn hai chiều $\mathbf{F}_l$ cho chuỗi văn bản s được xác định.

2.3. Hợp nhất và hiệu chỉnh lỗi

Phương pháp dự đoán song song của Transformer sử dụng đầu vào nhiễu, thường là các xấp xỉ từ dự đoán hình ảnh hoặc đặc trưng hình ảnh. Chẳng hạn, như ví dụ được thể hiện trong hình 2, ký tự kỳ vọng ở vị trí thứ hai là 'G', ở vị trí thứ ba là 'H'. Tuy nhiên, có thể do ảnh bị mờ, đứt nét và che khuất, ký tự thực tế nhận được là '6' và 'N'. Khi đó, '6' và 'N' trở thành nhiễu và làm giảm độ tin cậy của dự đoán. Mô hình hợp nhất và hiệu chỉnh lỗi lặp (hình 1) được đề xuất sử dụng để khắc phục vấn đề này. Cụ thể, đối với mỗi chuỗi đầu vào $\boldsymbol{s} = (s_1, \dots, s_n)$, mô hình ngôn ngữ sẽ được thực hiện lặp đi lặp lại M lần với các phép gán khác nhau cho s. Đối với lần lặp đầu tiên ($s_{i=1}$) là dự đoán xác suất bởi mô hình dự đoán chuỗi ký tự từ hình ảnh. Đối với các lần lặp tiếp theo ($s_{i\geq 2}$) là dự đoán xác suất từ mô hình hợp nhất trong lần lặp trước. Bằng cách này, mô hình có thể sửa chữa dự đoán hình ảnh một cách lặp đi lặp lại.

Một quan sát khác là các phương pháp dựa trên Transformer thường gặp vấn đề không cân đối về chiều dài hay nói cách khác là chiều dài văn bản được dự đoán không đúng với thực tế. Điều này dẫn tới khó khăn trong việc hiệu chỉnh. Trong cách tiếp cận của chúng tôi, các đặc trưng hình ảnh và các đặc trưng ngôn ngữ được hợp nhất nhiều lần. Trong quá trình hợp nhất, độ dài văn bản được dự đoán cũng được tinh chỉnh dần dần. Như vậy, cơ chế hợp nhất ở đây ngoài mục đích kết hợp các thông tin ngữ cảnh từ mô hình ngôn ngữ để hiệu chỉnh, tăng độ chính xác dự đoán còn có vai trò điều chỉnh chiều dài chuỗi được dự đoán trong quá trình lặp.

Để căn chỉnh các đặc trưng ngôn ngữ và đặc trưng hình ảnh, cơ chế cổng (gated mechanism) [14] được sử dụng để đưa ra quyết định cuối cùng:

$$\mathbf{FM} = \sigma\big([\mathbf{F}_v, \mathbf{F}_l]\mathbf{W}_f\big) \tag{10}$$

$$\mathbf{F}_f = \mathbf{FM} \odot \mathbf{F}_v + (1 - \mathbf{FM}) \odot \mathbf{F}_l \tag{11}$$

Trong đó, $\mathbf{W}_f \in \mathbb{R}^{2d\times d}$ là một tham số được khởi tạo; $\mathbf{F}_v, \mathbf{F}_l, \mathbf{F}_f$ lần lượt là các đặc trưng hình ảnh, đặc trưng ngôn ngữ và đặc trưng hợp nhất; σ là hàm sigmoid; $\mathbf{FM} \in \mathbb{R}^{t\times d}$ tự động lựa chọn các đặc trưng từ $\mathbf{F}_v$ và $\mathbf{F}_l$.

2.4. Huấn luyện mô hình

Để tăng tính hiệu quả, mô hình được huấn luyện liền mạch (end-to-end) với các mục tiêu đa nhiệm sau:

$$\mathcal{L} = \lambda_v \mathcal{L}_v + \frac{\lambda_l}{M}\sum_{i=1}^{M} \mathcal{L}_l^i + \frac{1}{M}\sum_{i=1}^{M} \mathcal{L}_f^i \tag{12}$$

Trong đó, $\mathcal{L}_v$, $\mathcal{L}_l$ và $\mathcal{L}_f$ là các chi phí entropy chéo (cross entropy losses) lần lượt từ $\mathbf{F}_v$, $\mathbf{F}_l$ và $\mathbf{F}_f$; $\mathcal{L}_l^i$ và $\mathcal{L}_f^i$ là các chi phí tại lần lặp thứ i; λ_v và λ_l là các hệ số cân bằng; M là số lần lặp.

Bên cạnh đó, để phát huy lợi thế của cơ chế tự điều khiển (autonomous) và hiệu chỉnh lặp, phương pháp huấn luyện bán giám sát dựa trên cơ chế tự học với tập hợp các dự đoán lặp được đề xuất áp dụng. Ý tưởng cơ bản của tự huấn luyện như sau: Trước tiên, tạo ra các nhãn giả bằng chính mô hình, sau đó huấn luyện lại mô hình bằng cách sử dụng các nhãn giả bổ sung. Như vậy, vấn đề chính nằm ở việc xây dựng các nhãn giả chất lượng cao.

Để lọc các nhãn giả nhiễu, chúng tôi dựa trên các quy tắc sau đây: i) Độ tin cậy tối thiểu của các ký tự trong một vùng văn bản được coi là độ chắc chắn của văn bản (text certainty). ii) Các

dự đoán lặp lại của mỗi ký tự được coi như một tập hợp để làm mờ tác động của các nhãn nhiễu. Hàm lọc được định nghĩa:

$$\begin{cases} \mathcal{C} & = \min\limits_{1\le t\le T} e^{\mathbb{E}[\log P(s_t)]} \\ P(s_t) & = \max\limits_{1\le m\le M} P_m(s_t) \end{cases} \quad (13)$$

Trong đó, $\mathcal{C}$ là sự chắc chắn tối thiểu của một vùng văn bản. $P_m(s_t)$ là phân bố xác suất của ký tự thứ t tại lần lặp thứ m. Quy trình huấn luyện được mô tả trong Thuật toán 1, trong đó T là một giá trị ngưỡng. B_l, B_u là các batch huấn luyện từ dữ liệu có nhãn và không có nhãn, *max* là số bước huấn luyện tối đa và *n* là số bước để cập nhật nhãn giả.

Thuật toán 1: Tự học kết hợp (Ensemble Self-training)

Đầu vào:

+ Tập ảnh đã được gán nhãn: X

+ Tập nhãn: Y (chứa toàn bộ nhãn đã gán cho tập X)

+ Tập các ảnh chưa gán nhãn: U

Đầu ra: Mô hình được huấn luyện

1. Huấn luyện các tham số θ_0 của mô hình với (X,Y) sử dụng công thức (12).
2. Sử dụng θ_0 để sinh tập nhãn giả V cho U
3. Lấy (U′, V′) bằng cách lọc (U, V) với C < T sử dụng công thức (13)
4. for i = 1, . . . , *max* do
5. if i == n then
6. Cập nhật V sử dụng θ_i
7. Lấy (U′, V′) bằng cách lọc (U, V) với C < T sử dụng công thức (13)
8. end if
9. Lấy mẫu $B_l = (\mathcal{X}_b, \mathcal{Y}_b) \subsetneq (\mathcal{X}, \mathcal{Y}), B_u = (\mathcal{U}'_b, \mathcal{V}'_b) \subsetneq (\mathcal{U}', \mathcal{V}')$
10. Cập nhật θi với B_l, B_u sử dụng công thức (12).
11. end for

3. KẾT QUẢ VÀ THẢO LUẬN

Nhóm tác giả sử dụng môi trường python 3 để cài đặt thuật toán và mô hình thử nghiệm. Cấu hình máy chạy thử Intel core i7-9700 CPU 4.7 GHz (Max Turbo Frequency), 32 GB RAM. Máy tính được trang bị Card đồ họa VGA Nvidia Tesla K80 12 GB × 2 GDDR5. Hiệu quả của thuật toán được đánh giá trên 02 tập dữ liệu thử nghiệm:

• **VinText** [15]: Là tập dữ liệu văn bản ngoại cảnh tiếng Việt VinText được thu thập và công bố bởi Viện VinAI, với tổng số 1200 ảnh cho training và 300 ảnh cho testing. Các ảnh trong tập dữ liệu này được thu thập ở khu vực Hà Nội và các vùng lân cận, rất đa dạng, gồm ảnh chụp biển quảng cáo, biển tên đường, tên cửa hiệu, văn phòng, dòng chữ trên các phương tiện giao thông, v.v. Các ảnh trong tập dữ liệu này được gán nhãn mức từ.

• **VNSceneText**: Là tập ảnh văn bản ngoại cảnh do nhóm tác giả tự thu thập trực tiếp trên đường phố khu vực Bà Rịa – Vũng Tàu và thành phố Hồ Chí Minh trong điều kiện hoàn toàn tự nhiên bằng các thiết bị smart phone (Iphone 7, Iphone 12, Oppo Reno 8). Tập dữ liệu VNSceneText bao gồm tổng số 3000 ảnh, trong đó 2400 ảnh dùng cho training và 600 ảnh dùng cho testing. Văn bản trong tập dữ liệu này rất đa dạng về chủng loại, điển hình gồm ảnh chụp từ biển quảng cáo, biển chỉ dẫn giao thông, tên đường phố, biển hiệu trên các tòa nhà, phương tiện giao thông và ảnh chụp từ nhiều loại văn bản giấy tờ khác. Để huấn luyện mô hình, chúng tôi đã

tiến hành gán nhãn mức từ (word-level-annotations). Để có dữ liệu gán nhãn, trước tiên chúng tôi đã tiến hành phân đoạn các ảnh từ (word images) trên mỗi ảnh văn bản đầu vào. Toàn bộ tập từ cắt được sau đó sẽ được duyệt lại để loại bỏ những ảnh kém chất lượng. Tập dữ liệu thu được sau bước lọc bao gồm 97.500 ảnh từ (word images) đảm bảo đa dạng về chất lượng, kiểu dáng, kích thước, font chữ.

3.1. Độ đo đánh giá

Hiệu quả của phương pháp được đánh giá dựa trên hai độ đo chính là độ chính xác nhận dạng và thời gian thực hiện. Đối với một hệ thống OCR nói chung, độ chính xác nhận dạng thường được đo bằng độ chính xác mức từ (Word_Acc) và độ chính xác mức ký tự (Char_Acc). Trong đó:

• *Độ chính xác nhận dạng mức ký tự (Char_Acc)*: được tính bằng số ký tự nhận dạng đúng (NC_T) trên tổng số ký tự cần nhận dạng (NC):

$$Char\_Acc = \frac{NC_T}{NC} \tag{14}$$

• *Độ chính xác mức từ (Word_Acc)*: Được tính bằng số từ nhận dạng đúng (NW_T) trên tổng số từ cần nhận dạng (NW):

$$Word\_Acc = \frac{NW_T}{NW} \tag{15}$$

Trong đó, một từ được coi là nhận dạng đúng nếu các ký tự trong nó đều nhận dạng đúng.

3.2. Kết quả thực nghiệm

Hiệu quả của phương pháp được đánh giá trên hai tập dữ liệu kiểm thử (testing data) tương ứng của VinText (300 ảnh) và VNSceneText (600 ảnh). Với cách chuẩn bị dữ liệu tương tự như như công đoạn training mô hình: Các ảnh trong hai tập dữ liệu kiểm thử trước tiên được đưa qua bộ phát hiện văn bản để tìm kiếm và tách ra các ảnh từ (word images), sau đó, tiến hành gán nhãn mức từ (word image annotation). Thông tin của hai tập dữ liệu này được mô tả cụ thể trên bảng 1.

Bảng 1. *Dữ liệu thử nghiệm.*

Tập dữ liệu	***Số từ (NW)***	***Số ký tự (NC)***
VinText	12700	79375
VNSceneText	28950	196700

Do đặc điểm cấu trúc giống nhau nên một số cặp ký tự hoa/thường sau đây được quy về cùng một lớp mẫu: C/c, O/o, Ô/ô, Ồ/ồ, Ổ/ổ, Ỗ/ỗ, Ố/ố, Ộ/ộ, Ơ/ơ, Ờ/ờ, Ở/ở, Ỡ/ỡ, Ớ/ớ, Ớ/ợ, S/s, V/v. Kết quả thực nghiệm của phương pháp phát hiện văn bản ngoại cảnh từ ảnh được cho trong bảng 2.

Bảng 2. *Kết quả thực nghiệm.*

Tập dữ liệu	***NW***	***NC***	***NW_T***	***NC_T***	***Word_Acc (%)***	***Char_Acc (%)***
VinText	12700	79375	10398	73835	81.87	93.02
VNSceneText	28950	196700	23901	185547	82.56	94.33

Các kết quả thực nghiệm đã chỉ ra phương pháp đề xuất đạt độ chính xác mức từ và mức ký tự lần lượt đối với tập dữ liệu VinText là 81.87% và 93.02%, đối với tập VNSceneText là 82.56% và 94.33%. Các kết quả trên bảng 2 được đánh giá trên tập các tập dữ liệu đầu vào sau khi đã loại bỏ nhiễu và những ảnh có chất lượng thấp. Do vậy, khi áp dụng trong thực tế, độ chính xác của thuật toán có thể bị giảm trên những ảnh đầu vào có chất lượng thấp hoặc chứa nhiều bất thường/biến dạng. Đối với những trường hợp này, chúng ta có thể áp dụng các kỹ thuật tiền xử lý, nắn chỉnh và tăng cường chất lượng hình ảnh trước khi thực hiện các thuật toán phát hiện và nhận dạng văn bản trong ảnh.

Để đánh giá hiệu quả về mặt thời gian, nhóm đã tiến hành ước lượng thời gian thực thi của phương pháp trên lần lượt từng tập dữ liệu. Kết quả thực nghiệm cho thấy, thời gian nhận dạng

trung bình đối với mỗi ảnh trong tập dữ liệu VinText mất khoảng 0.15 giây, và trên tập VNSceneText là 0.13s.

a) Ảnh đầu vào *b) Phát hiện văn bản trên ảnh* *c) Kết quả nhận dạng*

Hình 3. *Kết quả nhận dạng văn bản trong ảnh ngoại cảnh.*

Bên cạnh đó, phân tích kỹ hơn các kết quả thực nghiệm cho thấy, phương pháp đề xuất đạt hiệu quả tốt trên nhiều loại văn bản trong nhiều ngữ cảnh khác nhau, có khả năng lưu giữ được ngữ nghĩa văn bản khi cố gắng nhận dạng các ký tự khó. Ngoài ra, để huấn luyện thuật toán, chúng ta chỉ cần gán nhãn dữ liệu ở mức từ (word-level annotation) mà không cần đến dữ liệu huấn luyện dùng để tạo mô hình phân loại kí tự. Tuy nhiên, phương pháp cũng còn một số hạn chế như: Hiệu suất chưa cao khi đối mặt với các kiểu chữ viết tay hoặc ngôn ngữ không chuẩn.

4. KẾT LUẬN

Trong bài báo này, chúng tôi đề xuất một mô hình hiệu quả để giải quyết bài toán nhận dạng văn bản tiếng Việt trong ảnh ngoại cảnh dựa trên nền tảng học sâu. Mô hình nhận dạng được đề xuất dựa trên ý tưởng kết hợp ba luồng xử lý đồng thời trong một công đoạn nhận dạng, bao gồm: Nhận dạng (dự đoán) chuỗi ký tự từ hình ảnh; Xử lý ngữ cảnh; Hợp nhất và hiệu chỉnh chỉnh lỗi. Các kết quả thực nghiệm cho thấy phương pháp đề xuất đạt hiệu quả tốt trên nhiều loại văn bản trong nhiều ngữ cảnh khác nhau, có khả năng lưu giữ được ngữ nghĩa văn bản khi cố gắng nhận dạng các ký tự khó. Để huấn luyện thuật toán, chúng ta chỉ cần gán nhãn dữ liệu ở mức từ (word-level annotation) mà không cần đến dữ liệu huấn luyện dùng để tạo mô hình phân loại kí tự. Tuy nhiên, phương pháp cũng còn một số hạn chế như: Hiệu suất chưa cao khi đối mặt với các kiểu chữ viết tay hoặc ngôn ngữ không chuẩn. Trong thời gian tới, nhóm tác giả sẽ tiếp tục cải thiện thuật toán để tăng khả năng thích nghi và đối phó với các trường hợp văn bản đầu vào bất thường, có hình dạng phức tạp.

Lời cảm ơn: *Nhóm tác giả xin chân thành cảm ơn nhiệm vụ cao cấp, mã số NVCC02.01/23-23 đã hỗ trợ trong quá trình thực hiện nghiên cứu này.*

TÀI LIỆU THAM KHẢO

[1]. B. Shi, X. Bai, and C. Yao, *"An end-to-end trainable neural network for image-based sequence recognition and its application to scene text"*.

[2]. W. Liu, C. Chen, K.-Y. K. Wong, Z. Su, and J. Han, *"STAR-Net: A spatial attention residue network for scene text recognition,"* in Proc. Brit. Mach. Vision Conf. (BMVC). BMVA Press, pp. 43.1–43.13, (2016).

[3]. W. Liu, C. Chen, and K.-Y. K. Wong, *"Char-net: A characteraware neural network for distorted scene text recognition,"* in Proc. AAAI Conf. on Artif. Intell., (2018).

[4]. P. He, W. Huang, Y. Qiao, C. C. Loy, and X. Tang, *"Reading scene text in deep convolutional sequences,"* in Proc. AAAI Conf. on Artif. Intell., (2016).

[5]. F. Borisyuk, A. Gordo, and V. Sivakumar, *"Rosetta: Large scale system for text detection and recognition in images,"* in Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, pp. 71–79, (2018).

[6]. C.-Y. Lee and S. Osindero, *"Recursive recurrent nets with attention modeling for OCR in the wild,"* in Proc. IEEE Conf. on Comp. Vision and Pattern Recognit., pp. 2231–2239, (2016).

[7]. J. Wang and X. Hu, *"Gated recurrent convolution neural network for OCR,"* in Proc. Adv. in Neural

Inf. Process. Syst., pp. 335–344, (2017).
[8]. Y. Liu, Z. Wang, H. Jin, and I. Wassell, *"Synthetically supervised feature learning for scene text recognition,"* in Proc. Eur. Conf. on Comp. Vision (ECCV), pp. 435–451, (2018).
[9]. M. Liao, J. Zhang, Z. Wan, F. Xie, J. Liang, P. Lyu, C. Yao, and X. Bai, *"Scene text recognition from two-dimensional perspective,"* ArXiv, vol. abs/1809.06508, (2018).
[10]. Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun, *"Deep Residual Learning for Image Recognition"*, https://doi.org/10.48550/arXiv.1512.03385, (2015).
[11]. Z. Wan, F. Xie, Y. Liu, X. Bai, and C. Yao, *"2D-CTC for scene text recognition,"* (2019) .
[12]. K. He, X. Zhang, S. Ren, and J. Sun, *"Deep residual learning for image recognition,"* Proc. IEEE Conf. on Comp. Vision and Pattern Recognit. (CVPR), pp. 770–778, (2015).
[13]. F. Yin, Y.-C. Wu, X.-Y. Zhang, and C.-L. Liu, *"Scene text recognition with sliding convolutional character models,"* arXiv preprint arXiv:1709.01727, (2017).
[14]. Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, Yongdong Zhang, *"Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition"*, 2021 arXiv:2103.06495, https://doi.org/10.48550/arXiv.2103.06495, (2021).
[15]. N. Nguyen et al., *"Dictionary-guided Scene Text Recognition,"* 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 7379-7388, doi: 10.1109/CVPR46437.2021.00730.
[16]. N. T. Pham, V. D. Pham, Q. Nguyen-Van, B. H. Nguyen, D. N. Minh Dang and S. D. Nguyen, *"Vietnamese Scene Text Detection and Recognition using Deep Learning: An Empirical Study,"* 6th International Conference on Green Technology and Sustainable Development (GTSD), Nha Trang City, Vietnam, pp. 213-218, (2022), doi: 10.1109/GTSD54989.2022.9989248
[17]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, *"Attention Is All You Need"*, https://doi.org/10.48550/arXiv.1706.03762, 2023.
[18]. Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, Jianming Liang, *"UNet++: A Nested U-Net Architecture for Medical Image Segmentation"*, https://doi.org/10.48550/arXiv.1807.10165, 2018.

ABSTRACT

Vietnamese text recognition in scene images using deep learning

This article proposes an effective method for recognizing Vietnamese text in scene images. The proposed method is based on the idea of combining three processing tasks simultaneously in one recognition stage, including (i) Recognizing (predicting) character sequences from images; (ii) Context processing; and (iii) Fusing and iterative correction. The effectiveness of this method was carried out on two Vietnamese scene image datasets collected from reality: VinText and VnSceneText. Experimental results show that the proposed method is capable of detecting text of any shape and size with high and stable accuracy. Specifically, the method achieves word-level accuracy, character-level accuracy is (81.87%, 93.02%) and (82.56%, 94.33%) for the test datasets, respectively.

Keywords: Detection; Recognition; Feature; Probability; Accuracy.