**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN 1**



**ĐỒ ÁN TỐT NGHIỆP**

**ĐỀ TÀI**

**NHẬN DIỆN CHỮ TIẾNG VIỆT VÀ ỨNG DỤNG VÀO THIẾT BỊ ĐỌC SÁCH CHO NGƯỜI KHIẾM THỊ TẠI THƯ VIỆN**

|  |  |  |
| --- | --- | --- |
| **Giảng viên hướng dẫn:** | **TS. NGUYỄN TRỌNG KHÁNH** | |
| **Sinh viên thực hiện:** | | **NGUYỄN VŨ CHIẾN** | |
| **Mã sinh viên:** | | **B17DCCN085** | |
| **Lớp:** | | **E17CN01** | |
| **Khoá:** | | **2017 - 2022** | |
| **Hệ:** | | **ĐẠI HỌC CHÍNH QUY** | |

**Hà Nội – 12/2021**

**Hà Nội, tháng 11 năm 2017**

***Hà Nội, 12/2015***

**C VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN 1**



**ĐỒ ÁN TỐT NGHIỆP**

**ĐỀ TÀI**

**NGHIÊN CỨU PHƯƠNG PHÁP NHẬN DẠNG KÝ TỰ QUANG HỌC VÀ XÂY DỰNG HỆ THỐNG NHẬN DẠNG CHỨNG MINH THƯ NHÂN DÂN**

|  |  |  |
| --- | --- | --- |
| **Giảng viên hướng dẫn:** | **TS. NGUYỄN TRỌNG KHÁNH** | |
| **Sinh viên thực hiện:** | | **CHU VĂN AN** | |
| **Mã sinh viên:** | | **B16DCCN001** | |
| **Lớp:** | | **D16CNPM1** | |
| **Khoá:** | | **2016 - 2021** | |
| **Hệ:** | | **ĐẠI HỌC CHÍNH QUY** | |

**Hà Nội, tháng 11 năm 2017**

***Hà Nội, 12/2015***

**C VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN 1**



**ĐỒ ÁN TỐT NGHIỆP**

**ĐỀ TÀI**

**NGHIÊN CỨU PHƯƠNG PHÁP NHẬN DẠNG KÍ TỰ QUANG HỌC VÀ XÂY DỰNG HỆ THỐNG NHẬN DẠNG CHỨNG MINH THƯ NHÂN DÂN**

|  |  |  |
| --- | --- | --- |
| **Giảng viên hướng dẫn:** | **TS. NGUYỄN TRỌNG KHÁNH** | |
| **Sinh viên thực hiện:** | | **CHU VĂN AN** | |
| **Mã sinh viên:** | | **B16DCCN001** | |
| **Lớp:** | | **D16CNPM1** | |
| **Khoá:** | | **2016 - 2021** | |
| **Hệ:** | | **ĐẠI HỌC CHÍNH QUY** | |

**Hà Nội, tháng 11 năm 2017**

***Hà Nội, 12/2015***

**C VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN 1**



**ĐỒ ÁN TỐT NGHIỆP**

**ĐỀ TÀI**

**NGHIÊN CỨU PHƯƠNG PHÁP NHẬN DẠNG KÝ TỰ QUANG HỌC VÀ XÂY DỰNG HỆ THỐNG NHẬN DẠNG CHỨNG MINH THƯ NHÂN DÂN**

|  |  |  |
| --- | --- | --- |
| **Giảng viên hướng dẫn:** | **TS. NGUYỄN TRỌNG KHÁNH** | |
| **Sinh viên thực hiện:** | | **CHU VĂN AN** | |
| **Mã sinh viên:** | | **B16DCCN001** | |
| **Lớp:** | | **D16CNPM1** | |
| **Khoá:** | | **2016 - 2021** | |
| **Hệ:** | | **ĐẠI HỌC CHÍNH QUY** | |

**Hà Nội, tháng 11 năm 2017**

***Hà Nội, 12/2015***

# LỜI CẢM ƠN

Lời đầu tiên em xin cảm ơn thầy Nguyễn Trọng Khánh vì sự định hướng và giúp đỡ tận tình của thầy trong quá trình thực hiện đồ án. Nhờ sự giúp đỡ của thầy, em đã hiểu biết thêm nhiều hơn về lĩnh vực học máy và xử lý ảnh để thực hiện đồ án này. Được học tập và làm việc với Thầy đã giúp em tích lũy thêm được nhiều kiến thức và kinh nghiệm quý báu trong lĩnh vực mà em đang nghiên cứu.

Em cũng xin được gửi lời cảm ơn tới các thầy cô trong Học viện và trong Khoa Công nghệ thông tin I đã truyền tải rất nhiều kiến thức bổ ích với ngành học mà em đã lựa chọn. Em chúc các thầy cô luôn có sức khỏe, hạnh phúc và có thêm thật nhiều nhiệt huyết để vững bước chèo lái con thuyền trồng người đến bờ bến thành công.

Đồng thời em cũng muốn gửi lời cảm ơn tới gia đình của mình. Sự tin tưởng và tình yêu thương từ gia đình chính là động lực lớn lao nhất đối với em để vượt qua những khó khăn và trở ngại.

Lời cảm ơn cuối cùng, em xin gửi tới những người bạn, những người chiến hữu đã đồng hành và giúp đỡ em trong suốt quá trình học tập và làm đồ án. Nhờ có các bạn mà quãng đời sinh viên của em đã trở thành quãng thời gian đáng nhớ và học hỏi được nhiều điều và em sẽ mãi không quên được khoảng thời gian này.

Em chúc tất cả mọi người thật nhiều sức khỏe và thành công!

Hà Nội, tháng 12 năm 2021

**Sinh viên**

**Nguyễn Vũ Chiến**

**NHẬN XÉT, ĐÁNH GIÁ, CHO ĐIỂM**

**(Của giảng viên hướng dẫn)**

**Điểm: (bằng chữ: )**

**Đồng ý/Không đồng ý** cho sinh viên bảo vệ trước hội đồng chấm tốt nghiệp.

Hà Nội, tháng 12 năm 2021

**CÁN BỘ - GIẢNG VIÊN HƯỚNG DẪN**

**NGUYỄN TRỌNG KHÁNH**

**NHẬN XÉT, ĐÁNH GIÁ, CHO ĐIỂM**

**(Của giảng viên phản biện)**

**Điểm: (bằng chữ: )**

**Đồng ý/Không đồng ý** cho sinh viên bảo vệ trước hội đồng chấm tốt nghiệp.

Hà Nội, tháng 12 năm 2021

**CÁN BỘ - GIẢNG VIÊN PHẢN BIỆN**

# MỤC LỤC

[LỜI CẢM ƠN 2](#_Toc91702066)

[MỤC LỤC 5](#_Toc91702067)

[DANH MỤC HÌNH ẢNH 7](#_Toc91702068)

[DANH MỤC CÁC BẢNG 9](#_Toc91702069)

[DANH MỤC TỪ VIẾT TẮT 10](#_Toc91702070)

[LỜI MỞ ĐẦU 11](#_Toc91702071)

[CHƯƠNG 1: GIỚI THIỆU CHUNG 12](#_Toc91702072)

[1. Ứng dụng học sâu trong hỗ trợ người khiếm thị tiếp cận sách 12](#_Toc91702073)

[2. Mục tiêu và phạm vi nghiên cứu 12](#_Toc91702074)

[3. Một số nghiên cứu liên quan 13](#_Toc91702075)

[4. Kết luận 14](#_Toc91702076)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ CÔNG NGHỆ 15](#_Toc91702077)

[1. NỀN TÀNG CHUNG VỀ CÔNG NGHỆ 15](#_Toc91702078)

[1.1 Giới thiệu bài toán nhận dạng kí tự quang học 15](#_Toc91702079)

[1.2. Tổng quan các bước nhận dạng kí tự quang học 17](#_Toc91702080)

[1.3. Các phương pháp chuyển văn bản thành âm thanh 17](#_Toc91702081)

[1.4. Mục tiêu và phạm vi đồ án 19](#_Toc91702082)

[2. NHẬN DẠNG VĂN BẢN TIẾNG VIỆT 20](#_Toc91702083)

[2.1. Các phương pháp nhận dạng trích xuất thông tin từ ảnh 20](#_Toc91702084)

[2.1.2. Các phương pháp phát hiện vùng kí tự 24](#_Toc91702085)

[2.1.3. Các phương pháp nhận dạng chuỗi kí tự 28](#_Toc91702086)

[2.2. Xây dựng mô hình 45](#_Toc91702087)

[2.3. Kết luận 46](#_Toc91702088)

[CHƯƠNG 3: ĐỀ XUẤT GIẢI PHÁP ỨNG DỤNG 47](#_Toc91702089)

[1. Kịch bản hệ thống 47](#_Toc91702090)

[2. Xây dựng hệ thống 47](#_Toc91702091)

[2.1. Phân tích hệ thống 47](#_Toc91702092)

[2.2. Thiết kế hệ thống 48](#_Toc91702093)

[3. Huấn luyện mô hình và nhận dạng trích xuất thông tin 49](#_Toc91702094)

[3.1. Huấn luyện mô hình 49](#_Toc91702095)

[3.2. Nhận dạng trích xuất thông tin 50](#_Toc91702096)

[4. Thử nghiệm hệ thống 51](#_Toc91702097)

[4.1. Cài đặt huấn luyện các mô hình 51](#_Toc91702098)

[4.2. Nhận dạng trích xuất thông tin 54](#_Toc91702099)

[5. Kết luận 62](#_Toc91702100)

[CHƯƠNG 4: KẾT LUẬN 63](#_Toc91702101)

[1. Đóng góp của đồ án 63](#_Toc91702102)

[2. Hướng phát triển tiếp theo 63](#_Toc91702103)

# DANH MỤC HÌNH ẢNH

[Hình 1. Hệ thống Kurzweil 1000 với máy quét hình ảnh 14](#_Toc91871988)

[Hình 2. Thiết bị OrCam MyEye 15](#_Toc91871989)

[Hình 3. Công nghệ OCR giúp đưa tài liệu lên hệ thống và trích xuất thông tin nhanh chóng [25] 17](#_Toc91871990)

[Hình 4. Xác định xe vi phạm thông qua nhận diện biển số xe [26] 17](#_Toc91871991)

[Hình 5. Mô hình các thành phần cơ bản của một hệ thống TTS [35] 18](#_Toc91871992)

[Hình 6. Mô hình luồng dữ liệu từ văn bản thành giọng nói 19](#_Toc91871993)

[Hình 7. Phát hiện một đám cháy nhỏ ở khoảng 30 mét [36] 21](#_Toc91871994)

[Hình 8. Những yếu tố anht hưởng đến cách một hình ảnh của đối tượng được ghi lại [36] 22](#_Toc91871995)

[Hình 9. Sự khác nhau giữa nhận diện hình ảnh và phá hiện đối tượng [38] 23](#_Toc91871996)

[Hình 10. Kết quả của bài toán phân đoạn ngữ nghĩa [39] 24](#_Toc91871997)

[Hình 11. Kết quả của bài toán Phân đoạn cá thể [39] 24](#_Toc91871998)

[Hình 12. Các bước của kỹ thuật nhận diện bằng cửa sổ trượt [40] 26](#_Toc91871999)

[Hình 13. Kiến trúc mô hình CRAFT [41] 28](#_Toc91872000)

[Hình 14. Kiến trúc mạng VGG16BN [41] 29](#_Toc91872001)

[Hình 15. Mạng VGG16BN áp dụng trong CRAFT [41] 29](#_Toc91872002)

[Hình 16. Kiến trúc CRNN [42] 31](#_Toc91872003)

[Hình 17. Trường tiếp nhận. Mỗi vectơ trong chuỗi đặc trưng được trích xuất được liên kết với một trường tiếp nhận trên hình ảnh đầu vào và có thể được coi là vectơ đặc trưng của trường đó [42]. 32](#_Toc91872004)

[Hình 18. : Cấu trúc bên trong tế bào LSTM 33](#_Toc91872005)

[Hình 19. Mô hình hoạt động của LSTM [44] 34](#_Toc91872006)

[Hình 20. Cấu trúc của LSTM hai chiều sâu. 35](#_Toc91872007)

[Hình 21. So sánh hiệu quả của các mô hình STR hiện có và thực nghiệm trên các mô hình đó của nhóm tác giả [12] 38](#_Toc91872008)

[Hình 22. Kiến trúc của STN. Mạng cục bộ hóa định vị các điểm cơ sở C, bộ tạo lưới tạo ra lưới lấy mẫu P, bộ lấy mẫu V tạo hình ảnh chuẩn hóa I' từ hình ảnh đầu vào I và P [10] 39](#_Toc91872009)

[Hình 23. Tập điểm cơ sở và bộ chuyển đổi TPS [10] 40](#_Toc91872010)

[Hình 24. STN căn chỉnh một số hình ảnh chuỗi kí tự thường gặp [10] 41](#_Toc91872011)

[Hình 25. Kiến trúc của mạng cục bộ trong STN [12] 42](#_Toc91872012)

[Hình 26. Sự khác biệt giữa khối thông thường (trái) và khối phần dư (phải) [23] 43](#_Toc91872013)

[Hình 27. Mạng ResNet được sử dụng trong TBRA [12] 44](#_Toc91872014)

[Hình 28. Cơ chế seq2seq phổ biến (không có attention) [33] 45](#_Toc91872015)

[Hình 29. Cơ chế seq2seq có attention [33] 45](#_Toc91872016)

[Hình 30. Hệ thống chuyển đổi văn bản thành giọng nói 47](#_Toc91872017)

[Hình 31. Biểu đồ ca sử dụng trong hệ thống 48](#_Toc91872018)

[Hình 32. Kiến trúc chung của hệ thống 49](#_Toc91872019)

[Hình 33. Quy trình huấn luyện mô hình 50](#_Toc91872020)

[Hình 34. Quy trình nhận dạng chuyển văn bản thành giọng nói của hệ thống 51](#_Toc91872021)

[Hình 35. Tham số đào tạo TBRA 54](#_Toc91872022)

[Hình 36. Kết quả huấn luyện TBRA 54](#_Toc91872023)

[Hình 37. Biểu đồ thể hiện độ hội tụ của accuracy 55](#_Toc91872024)

[Hình 38. Ảnh đầu vào với kịch bản 1 57](#_Toc91872025)

[Hình 39. Ảnh đầu vào với kịch bản 2 59](#_Toc91872026)

[Hình 40. Ảnh đầu vào với kịch bản 3 61](#_Toc91872027)

# DANH MỤC CÁC BẢNG

[Bảng 1. Môi trường cài đặt thử nghiệm việc huấn luyện các mô hình 52](#_Toc91871938)

[Bảng 2. Cấu hình hệ thống thử nghiệm 55](#_Toc91871939)

[Bảng 3.Các thư viện chính được sử dụng trong máy chủ hệ thống 56](#_Toc91871940)

[Bảng 4. Công nghệ sử dụng trong máy chủ web và trình duyệt phía người dùng 56](#_Toc91871941)

[Bảng 5. Một số trường hợp nhận diện sai với kịch bản 1 58](#_Toc91871942)

[Bảng 6. Một số trường hợp nhận diện sai với kịch bản 2 60](#_Toc91871943)

[Bảng 7. Một số trường hợp nhận diện sai với kịch bản 2 61](#_Toc91871944)

[Bảng 8. Kết quả thử nghiệm hệ thống 62](#_Toc91871945)

# DANH MỤC TỪ VIẾT TẮT

|  |  |
| --- | --- |
| OCR: Optical Character Recognition | Nhận dạng kí tự quang học |
| AI: Artificial Intelligence | Trí tuệ nhân tạo |
| CV: Computer Vision | Thị giác máy tính |
| STR: Scene Text Recognition | Nhận dạng văn bản cảnh |
| CNN: Convolution Neural Network | Mạng nơ-ron tích chập |
| DCNN: Deep Convolutional Neural Cetwork | Mạng nơ-ron tích chập sâu |
| DL: Deep Learning | Học sâu |
| LSTM: Long Short Term Memory | Bộ nhớ dài – ngắn hạn |
| SW: Sliding Windows | Cửa sổ trượt |
| CC: Connected Components | Các thành phần kết nối |
| HOG: Histogram of oriented gradients | Biểu đồ độ dốc định hướng |
| STN: Spatial Transformer Networks | Mạng biến áp không gian |
|  |  |
|  |  |
|  |  |
|  |  |

# LỜI MỞ ĐẦU

Ngày nay, trí tuệ nhân tạo đang được ứng dụng rộng rãi và mang đến những giải pháp thiết thực trong các vấn để của đời sống xã hội, khiến cho cuộc sống của mỗi người trở nên dễ dàng, tiện lợi đặc biệt đối với người khuyết tật. Nhờ các ứng dụng trí tuệ nhận tạo mà những khó khăn, trở ngại đã dần được loại bỏ, giúp nâng cao chất lượng cuộc sống của họ. Trong đó, lĩnh vực thị giác máy tính đang được sử dụng như một giải pháp để thay thế con mắt của người khiếm thị, với các hệ thống trợ giúp nhận diện vật thế trong không gian để trợ giúp việc di chuyển hay các trợ lý ảo có thể trợ giúp sử dụng vật dụng trong nhà, hay số hoá các đầu sách dưới dạng sách nói – mà một trong những giải pháp đã và đang được phát triển cho việc số hoá là sử dụng phương pháp nhận dạng kí tự quang học (Optical Character Recognition - OCR), được hình thành từ lĩnh nghiên cứu về nhận dạng mẫu, trí tuệ nhân tạo và xử lý ảnh giúp trích xuất các văn bản trong hình ảnh thành văn bản số.

Với mong muốn tìm hiểu về phương pháp nhận dạng kí tự quang học cũng như đóng góp thêm vào kho ứng dụng về nhận dạng một hệ thống nhận dạng thiết thực, hữu ích. Đồ án thực hiện nghiên cứu các bước trong bài toán OCR và đề xuất giải pháp ứng dụng OCR trong nhận dạng văn bản tiếng Việt áp dụng trong các thư viện, giúp người khiếm thị có thể tiếp cận được nhiều đầu sách hơn bằng cách trực tiếp chuyển đổi sách giấy thành âm thanh.

Nội dung đồ án bao gồm 4 chương:

* Chương 1: Giới thiệu chung
* Chương 2: Cơ sở lý thuyết và công nghệ
* Chương 3: Đề xuất phương pháp
* Chương 4: Thử nghiệm và đanh giá

# CHƯƠNG 1: GIỚI THIỆU CHUNG

## **1. Ứng dụng học sâu trong hỗ trợ người khiếm thị tiếp cận sách**

Đối với người khiếm thị, việc tiếp nhận kiến thức luôn có một rào cản khi chỉ có thể thông qua âm thanh như nghe qua radio, tivi hay đọc những quyển sách chữ nổi. Việc đa phần thông tin đều ở dạng văn bản khiến cho những thông tin từ sách báo hay internet khó có thể được họ tự mình tiếp thu. Hiện nay, các giải pháp về công nghệ học sâu và trí tuệ nhân tạo cũng được ứng dụng mạnh mẽ để trợ giúp người khiếm thị khi sử dụng máy tính. Trong đó có các chương trình phần mềm cho phép người dùng khiếm thị đọc văn bản được hiển thị trên màn hình với bộ tổng hợp giọng nói hoặc màn hình chữ nổi. Người dùng gửi lệnh bằng cách nhấn các tổ hợp phím khác nhau trên bàn phím máy tính hoặc màn hình chữ nổi để hướng dẫn bộ tổng hợp giọng nói phải nói gì và âm thanh sẽ tự động phát khi có sự thay đổi nội dung trên màn hình. Một lệnh có thể hướng dẫn bộ tổng hợp đọc hoặc đánh vần một từ, đọc một dòng hoặc toàn màn hình văn bản, tìm một chuỗi văn bản trên màn hình. Ngoài ra, nó cho phép người dùng thực hiện các chức năng nâng cao hơn, chẳng hạn như định vị văn bản được hiển thị bằng một màu nhất định, đọc các phần được chỉ định trước của màn hình theo yêu cầu, đọc văn bản được đánh dấu và xác định lựa chọn đang hoạt động trong menu. Người dùng cũng có thể sử dụng trình kiểm tra chính tả trong trình xử lý văn bản hoặc đọc các ô của bảng tính bằng trình đọc màn hình. Tuy nhiên với những thư viện, nơi có lượng đầu sách lớn, đa dạng và nhiều kiến thức vẫn còn đang là một nguồn khó tiếp cận do việc chuyển toàn bộ sách thành sách nói tiêu tốn một khoản kinh phí cũng như thời gian rất lớn. Chình vì vậy, để có thể khắc phục được vấn đề trên, giải pháp có thể được sử dụng là một thiết bị có giá thành rẻ, có chức năng biến những cuốn sách giấy trong thư viện thành những quyển sách nói, dễ sử dụng cho cả nhân viên thư viện lẫn người khiếm thị. Giải pháp được đề xuất trong đồ án là sử dụng hai công nghệ chính để có thể tạo nên một thiết bị chuyển từ văn bản trên giấy thành giọng nói, bao gồm: công nghệ phát hiện ký tự dựa trên mạng VGG-16 và nhận dạng ký tự dựa trên kết hợp nhiều mạng neuron học sâu. Với chức năng quan trọng nhất của hệ thống là nhận diện đúng ký tự tiếng Việt. Một số yêu cầu cần đảm bảo bởi hệ thống khi vận hành thực tế như:

+ Giá thành chí phí thấp.

+ Đảm bảo ổn định để có thể sử dụng liên tục.

+ Đảm bảo toàn vẹn nội dung của sách, không bị thiếu hay sai thứ tự của các từ.

+ Thuận tiện để bảo trì, nâng cấp.

## **2. Mục tiêu và phạm vi nghiên cứu**

Đồ án nghiên cứu và đưa ra giải pháp cho bài toán nhận diện và đọc sách giấy cho người khiếm thị. Nội dung nghiên cứu tâp trung vào nghiên cứu các giải pháp nhận diện ký tự từ ảnh quét được bằng thiết bị chuyên dụng và chuyển văn bản dự đoán thành giọng nói. Đầu vào của bài toán được giả định là hình chụp của trang sách với điều kiện ảnh sáng đủ tốt và nền không quá phức tạp, mục tiêu hướng đến sau khi xây dựng hệ thống là có thể đạt hiệu năng cao trong nhận diện và chuyển đổi văn bản nhận diện thành âm thanh và thông qua API để phát thành giọng đọc.

## **3. Một số nghiên cứu liên quan**

**\* Kurzweil 1000**

Phần mềm đọc và quét từng đoạt giải thưởng giúp những người khiếm thị có thể truy cập văn bản điện tử hoặc bản in. Nó kết hợp các công nghệ máy đọc truyền thống như quét, xử lý hình ảnh và chuyển văn bản thành giọng nói với các công cụ giao tiếp và năng suất để dễ dàng và nâng cao trải nghiệm đọc, viết và học tập của người dùng. Phần mềm nói to văn bản bằng nhiều giọng đọc tự nhiên có thể được sửa đổi theo sở thích cá nhân.

Ngoài ra, phần mềm cung cấp cho người dùng khả năng viết và chỉnh sửa tài liệu, hoàn thành các biểu mẫu đơn giản một cách độc lập. Nó bao gồm các tính năng ghi chú, tóm tắt nội dung và lập dàn ý cho văn bản.



Hình 1. Hệ thống Kurzweil 1000 với máy quét hình ảnh

**\* OrCam MyEye**

OrCam MyEye là thiết bị chuyển văn bản thành giọng nói tiên tiến nhất hiện nay. Thiết bị thông minh khai thác sức mạnh của Thị giác nhân tạo (Artificial Vision) để hỗ trợ những người khiếm thị. Nó có thể đọc văn bản ngay lập tức từ bất kỳ bề mặt in hoặc kỹ thuật số nào. Một trong những thuộc tính hữu ích nhất của thiết bị là thiết bị không cần kết nối Internet để hoạt động.

Có hai cách mà người dùng có thể sử dụng thiết bị để đọc văn bản. Đầu tiên là tính năng đọc văn bản đơn giản. Tính năng này sẽ đọc văn bản từ đầu trang cho đến cuối trang. Tùy chọn thứ hai là sử dụng tính năng đọc văn bản thông minh. Tính năng này cho phép người dùng hướng dẫn bằng lời nói thiết bị để tìm thông tin cụ thể trong văn bản. Ví dụ, khi đọc hóa đơn tiền điện, bạn có thể hướng dẫn thiết bị tìm "tổng số tiền đến hạn". Video này trình bày cách hoạt động của tính năng đọc thông minh.



Hình 2. Thiết bị OrCam MyEye

## **4. Kết luận**

Chương 1 đã giới thiệu về bài toán nhận dạng kí tự quang học cũng như đề xuất mục tiêu xây dựng hệ thống nhận dạng ký tự tiếng Việt và những sản phẩm với nghiên cứu tương tự trên thị trường.

Trong chương tiếp theo, đồ án sẽ trình bày chi tiết hơn về cơ sở lý thuyết và các nghiên cứu hiện có cho từng bước trong bài toán nhận dạng và chuyển đổi văn bản tiếng Việt thành giọng nói.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ CÔNG NGHỆ

Chương này giới thiệu tổng quan về bài toán nhận dạng kí tự quang học cũng như các bước thực hiện của một bài toán nhận dạng kí tự quang học điển hình, đồng thời giới thiệu về bài toán nhận dạng văn bản tiếng Việt, một bài toán con trong nhận dạng kí tự quang học.

## **1. NỀN TÀNG CHUNG VỀ CÔNG NGHỆ**

## **1.1 Giới thiệu bài toán nhận dạng kí tự quang học**

Nhận dạng kí tự quang học là phương pháp giúp chuyển các hình ảnh của chữ viết tay hoặc chữ đánh máy thành các văn bản tài liệu. Nó là một trong những đề tài nghiên cứu hấp dẫn, có nhiều ứng dụng thực tiễn trong cuộc sống giảm thiểu thời gian công sức cho con người trong việc chuyển đổi ảnh scan, ảnh chụp tài liệu … thành dạng văn bản.

Hiện nay, các phần mềm có tích hợp ứng dụng hoặc phần mềm nhận diện ký tự đang được sử dụng phổ biến trong nhiều ngành nghề, đem lại hiệu quả công việc cao, tiết kiệm thời gian làm việc và giảm bớt nhân sự. Có thể kể đến trong đó là hành chính công và giao thông với nhiều ứng phổ biến.

\* Trong lĩnh vực hành chính công

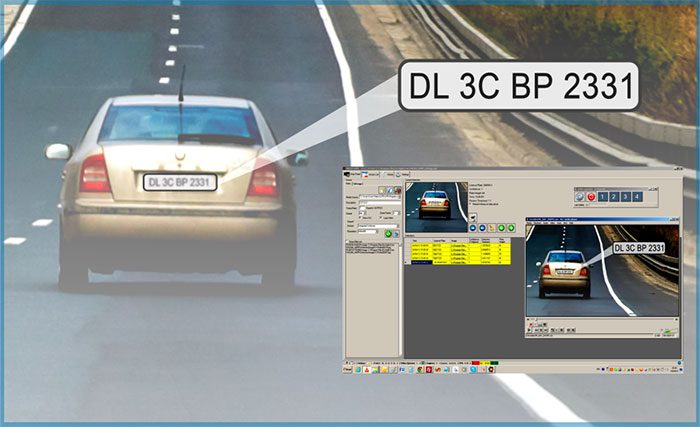
Công nghệ OCR đã mang lại rất nhiều tiện ích và rút ngắn thời gian xử lý các thủ tục hành chính. Bởi để phục vụ cho doanh nghiệp và người dân, các cơ quan cần tổng hợp và kiểm chứng nhiều danh mục thông tin khác nhau, cũng như bổ sung vào nguồn dữ liệu lớn để dần rút ngắn những thủ tục không cần thiết mỗi khi cần xử lý các yêu cầu pháp lý hướng đến cơ chế “một cửa”. Với sự kết với AI, OCR đã có thể nhận diện, trích xuất được các thông tin mong muốn từ những hồ sơ, văn bản được nộp vào hệ thống, từ đó giúp cho các thủ tục được thông suốt, đi đến đúng đơn vị cần xử lý. Đối với văn bản đánh máy, tỷ lệ chính xác của OCR lên đến hơn 98%. Tỷ lệ chính xác là rất cao tuy nhiên tỷ lệ nhỏ sai sót cũng có thể gây ra hậu quả lớn. Do đó, các cơ quan hành chính sử dụng công nghệ OCR như một phương tiện để hỗ trợ phân loại và xử lý văn bản đầu vào. Vậy nên, việc áp dụng công nghệ OCR vào hệ thống giúp quản lý được công việc hiệu quả, chuẩn hóa hoạt động của từng bộ phận, đơn giản hoá các quy trình, giảm gánh nặng, sai sót khi kiểm tra thủ tục hành chính, tăng năng suất lao động và tối ưu hoá chi phí cho doanh nghiệp, người dân.



Hình 3. Công nghệ OCR giúp đưa tài liệu lên hệ thống và trích xuất thông tin nhanh chóng [25]

\*Trong lĩnh vực giao thông

Ở các thành phố lớn hoặc trên các tuyến cao tốc, lưu lượng phương tiện giao thông ngày càng tăng tuy nhiên hệ thống giao thông hạ tầng lại chưa đáp ứng đủ gây ra tình trạng vi phạm luật lệ giao thông như chạy quá tốc độ, lấn làn, dừng đỗ sai quy định. Công nghệ OCR đã được áp dụng để hỗ trợ lực lượng chức năng có thể quản lý giao thông, xử phạt những hành vi vi phạm. Thông qua hệ thống camera giám sát, phần mềm nhận dạng kí tự quang học nhận diện biển số các phương tiện có hành vi vi phạm giao thông như vượt đèn đỏ, sai làn đường, không đội mũ bảo hiểm…Từ những hình ảnh này sau đó lấy thông tin của chủ phương tiện rồi tiến hành xử lý vi phạm. Ngoài ra công nghệ này còn hỗ trợ lưu trữ các hình ảnh, tìm hiểu diễn biến, nguyên nhân các vụ tai nạn.



Hình 4. Xác định xe vi phạm thông qua nhận diện biển số xe [26]

## **1.2. Tổng quan các bước nhận dạng kí tự quang học**

Một hệ thống nhận dạng kí tự quang học điển hình bao gồm các bước:

\* Thu thập hình ảnh: là quá trình chuyển hình ảnh thực của các đối tượng về dạng ảnh số, video số bằng cách sử dụng các thiết bị quay, chụp như máy ảnh số (camera), webcame hoặc các thiết bị smartphone. Đây thao tác cơ bản đầu tiên để cung cấp dữ liệu ảnh đầu vào cho các ứng dụng xử lý ảnh, xử lý video.

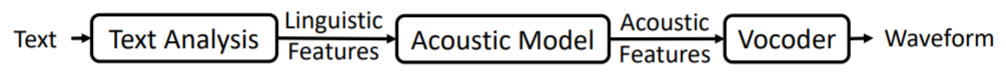
\* Tiền xử lý hình ảnh: là quá trình áp dụng các kĩ thuật xử lý nhằm phát hiện được đối tượng quan tâm trong ảnh và đồng thời xác định loại văn bản cần đọc để phục vụ cho công đoạn tiếp theo.

\* Phát hiện các vùng chứa kí tự trong ảnh: xác định các vùng chứa kí tự trong ảnh, các vùng có thể ở mức kí tự, từ hoặc dòng bằng các kĩ thuật xử lý ảnh cổ điển hoặc học máy.

\* Nhận dạng kí tự: là công đoạn xác định nhãn kí tự, từ hay dòng của hình ảnh được cắt từ bước phát hiện. Quá trình nhận dạng hiện nay thường áp dụng các mô hình học máy tổng hợp gồm một mô hình trích xuất các đặc trưng của ảnh và mô hình hồi quy để giúp nắm bắt thông tin ngữ cảnh của chuỗi kí tự.

## **1.3. Các phương pháp chuyển văn bản thành âm thanh**

Hiện nay, bài toán chuyển đổi văn bản thành giọng nói – Text to speech (TTS) hoặc tổng hợp giọng nói, nhằm mục đích tổng hợp văn bản đã cho thành giọng nói dễ hiểu và tự nhiên, là một chủ đề nghiên cứu nóng trong cộng đồng lời nói, ngôn ngữ và máy học và có ứng dụng rộng rãi trong ngành. Cũng như sự phát triển của học sâu (deep learning) và trí tuệ nhân tạo, TTS dựa trên mạng neuron đã cải thiện đáng kể chất lượng kết quả đầu ra của bài toán.



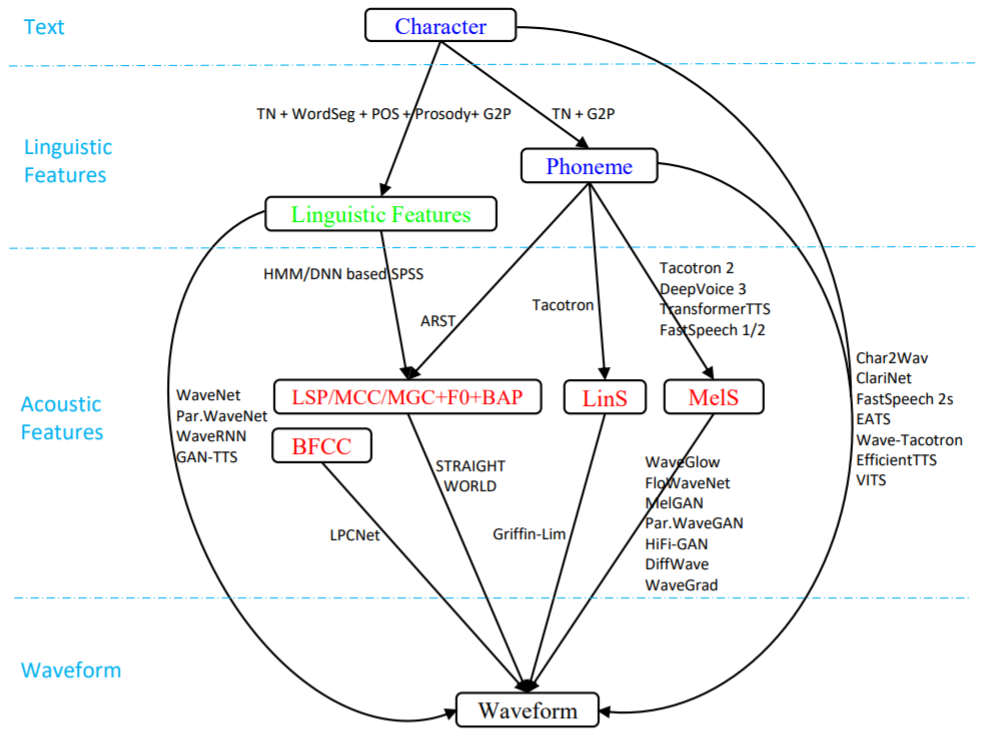
Hình 5. Mô hình các thành phần cơ bản của một hệ thống TTS [35]

Một hệ thống TTS hiện đại bao gồm ba thành phần cơ bản:

- Module phân tích văn bản (Text Analysis): chuyển đổi một chuỗi văn bản thành các đặc điểm ngôn ngữ.

- Mô hình âm thanh (Acoustic Model): tạo ra các đặc điểm âm thanh từ các đặc điểm ngôn ngữ.

- Bộ mã hóa (Vocoder): tổng hợp ra các dạng song âm thanh. từ các đặc điểm âm học.



Hình 6. Mô hình luồng dữ liệu từ văn bản thành giọng nói

Cách hoạt động của TTS theo luồng dữ liệu từ văn bản sang dạng song được tóm tắt như trong hình 3.2. Có một số biểu diễn dữ liệu trong quá trình chuyển đổi văn bản thành giọng nói:

1) Các ký tự (Text), là định dạng thô của văn bản.

2) Các đặc điểm của ngôn ngữ (Linguistic Feature), có được thông qua phân tích văn bản và chứa thông tin ngữ cảnh phong phú về cách phát âm và ngữ điệu. Ngoài ra có âm vị là một trong những yếu tố quan trọng nhất trong các đặc điểm ngôn ngữ và thường được sử dụng để biểu thị văn bản trong các mô hình TTS dựa trên mạng neuron.

3) Các tính năng âm học (Acoustic Feature), là các đại diện trừu tượng của dạng sóng giọng nói. Theo những thống kê tổng hợp về tham số đặc trưng cho giọng nói, cặp quang phổ vạch - LSP (line spectral pairs) [135], hệ số mel-cepstral - MCC (mel-cepstral coefficients), hệ số tổng quát mel MGC - (mel-generalized coefficients), F0 và các đặc điểm của dải BAP - (band aperiodicities) được sử dụng như các đặc trưng âm học, có thể dễ dàng chuyển đổi thành dạng sóng thông qua các bộ định tuyến như STRAIGHT và WORLD. Trong các mô hình TTS end- to-end dựa trên mạng neuron, phổ Mel (mel-spectrograms) hoặc phổ tuyến tính (linear-spectrograms) thường được sử dụng làm các đặc trưng âm thanh, được chuyển đổi thành dạng sóng bằng cách sử dụng các bộ mã hóa dựa trên mạng neuron.

4) Dạng sóng, định dạng cuối cùng của giọng nói. Có thể có các luồng dữ liệu khác nhau từ văn bản sang dạng sóng, bao gồm:

1) Ký tự → đặc điểm ngôn ngữ → đặc điểm âm thanh → dạng sóng;

2) Ký tự → âm vị → đặc điểm âm học → dạng sóng;

3) Ký tự → đặc điểm ngôn ngữ → dạng sóng;

4) Ký tự → âm vị → đặc điểm âm học → dạng sóng;

5) Ký tự → âm vị → dạng sóng, hoặc ký tự → dạng sóng.

## **1.4. Mục tiêu và phạm vi đồ án**

Mục tiêu chính của đồ án là nghiên cứu phương pháp nhận dạng kí tự quang học nói chung và áp dụng trong việc nhận dạng văn bản tiếng Việt cũng như xây dựng hệ thống chuyển đổi văn bản thành giọng nói hỗ trợ cho người khiếm thị trong các thư viện. Cụ thể đồ án sẽ:

- Về mặt lý thuyết, tìm hiểu về các phương pháp đang được sử dụng trong các bước của nhận dạng ký tự tiếng Việt bao gồm các mô hình phát hiện kí tự, mô hình nhận dạng kí tự và các kĩ thuật hậu xử lý.

- Về mặt ứng dụng, đồ án đặt mục tiêu xây dựng hệ thống có khả năng nhận dạng văn bản tiếng Việt từ trang sách sau đó thông tin được chuyển đổi thành giọng nói sử dụng API của Google.

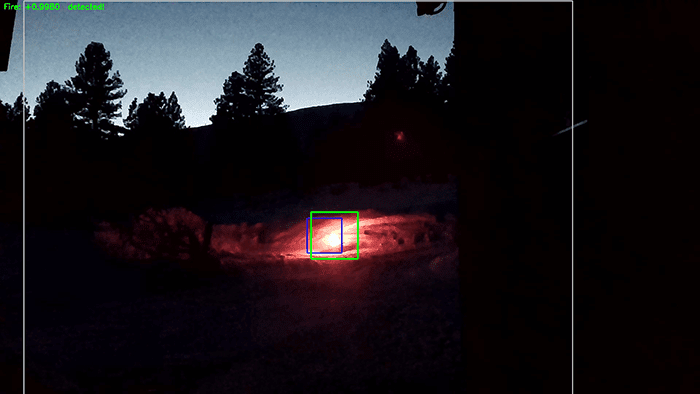
## **2. NHẬN DẠNG VĂN BẢN TIẾNG VIỆT**

## **2.1. Các phương pháp nhận dạng trích xuất thông tin từ ảnh**

Phần này sẽ giới thiệu về các phương pháp tiền xử lý, phát hiện vùng ký tự và nhận dạng ký tự hiện có và được áp dụng rộng rãi trong các bài toán nhận diện ký tự.

**2.1.1. Các phương pháp tiền xử lý hình ảnh**

Trước khi đi sâu vào từng phương pháp, ta cùng tìm hiểu về các nhóm bài toán trong thị giác máy tính (Computer Vision - CV). Bao gồm: phân lớp ảnh (Image classification), phân lớp và xác định vị trí (Classification with Localization), phát hiện đối tượng (Object detection), phân đoạn ảnh (Image segmentation).



Hình 7. Phát hiện một đám cháy nhỏ ở khoảng 30 mét [36]

**\* Phân lớp ảnh (Image classification):**

Mục tiêu của bài toán là phân tích hình ảnh đầu vào và trả về một nhãn phân loại hình ảnh. Nhãn luôn nằm trong một tập hợp các danh mục có thể được xác định trước. Hệ thống phân loại cũng có thể gán nhiều nhãn cho hình ảnh thông qua xác suất, chẳng hạn như {chó: 95%; mèo: 4%; gấu trúc: 1%}.

Kết quả bài toán phân lớp ảnh có thể áp dụng vào rất nhiều lĩnh vực như phân loại động vật, phân loại biển báo giao thông… Phân lớp ảnh cũng được cho là bài toán cơ sở cho một số bài toán khác trong CV. Trong thực tế bài toán này có khá nhiều thách thức:



Hình 8. Những yếu tố anht hưởng đến cách một hình ảnh của đối tượng được ghi lại [36]

* Đa dạng về góc nhìn (Viewpoint Variation): Đối tượng có thể được định hướng / xoay nhiều chiều phụ thuộc vào cách đối tượng được chụp.
* Đa dạng về tỉ lệ/ kích thước (Scale Variation): Với một chiêc áo có thể có nhiều size như S, M, L... Về mặt kỹ thuật, chúng đều giống nhau - một chiếc áo, nhưng chúng có kích thước khác nhau. Hơn nữa, cùng một chiếc áo size L sẽ trông khác biệt đáng kể khi nó được chụp gần so với khi nó được chụp từ xa hơn.
* Biến dạng (Deformation): Sự đa dạng hình ảnh của cùng một đối tượng có thể bị biến đổi theo các điều kiện khác nhau.
* Bị che khuất (Occlusion Variation): Đối tượng cần phân loại bị che khuất một phần.
* Điều kiện chiếu sáng (Illumination Variation): Cùng một ảnh chụp một vật nhưng ở những thiết lập chiếu sáng khác nhau có thể có những khác biệt đáng kể, gây sai sót trong quá trình phân loại.
* Ảnh hưởng bởi bối cảnh (Background Clutter): Những hình ảnh có "nhiễu" hoặc có rất nhiều đối tượng trong đó. Việc tìm kiếm một đối tượng cụ thể trong ảnh cũng gây khó khăn đối với con người. Nên sẽ khó khăn hơn nữa với máy tính không hiểu ngữ nghĩa của hình ảnh. Ngoài ra còn trường hợp đa dạng về biến thể trong một nhãn: Nhãn cần phân loại có rất nhiều loại ví dụ mèo thì có mèo tam thể, mèo mướp.

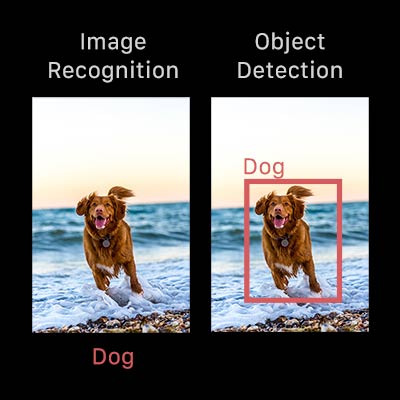
**\* Phân lớp và định vị (Classification with Localization)**

Trong bài toán phân lớp và định vị, một thuật toán được đào tạo có giám sát (supervised algorithm) để dự đoán lớp cũng như hộp giới hạn xung quanh đối tượng (bounding box) trong ảnh. Thuật ngữ ‘localization’ đề cập đến vị trí của đối tượng trong hình ảnh.[37]

**\* Phát hiện đối tượng (Object detection)**

Phát hiện đối tượng là một kỹ thuật thị giác máy tính để xác định và định vị các đối tượng trong một hình ảnh hoặc video. Cụ thể, phát hiện đối tượng vẽ các hộp giới hạn (bounding box) xung quanh các đối tượng được phát hiện này, cho phép xác định vị trí các đối tượng đang ở (hoặc cách chúng di chuyển qua, đối với đầu vào video) trong một cảnh nhất định.

Phát hiện đối tượng thường bị nhầm lẫn với nhận dạng hình ảnh. Nhận dạng hình ảnh gán nhãn cho hình ảnh. Hình ảnh một con chó được gắn nhãn "dog". Một bức ảnh của hai con chó, vẫn nhận được nhãn "dog". Mặt khác, phát hiện đối tượng, sẽ có một hộp xung quanh mỗi con chó và dán nhãn cho hộp đó là “dog”. Mô hình dự đoán vị trí của từng đối tượng và nhãn nào nên được áp dụng. Theo cách đó, phát hiện đối tượng cung cấp nhiều thông tin về hình ảnh hơn là nhận dạng. [38]



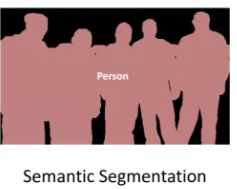
Hình 9. Sự khác nhau giữa nhận diện hình ảnh và phá hiện đối tượng [38]

**\* Phân đoạn ảnh (Image segmentation).**

Bài toán sẽ phân chia một hình ảnh thành nhiều vùng ảnh khác nhau. Image Segmentation cũng có mục tiêu giống như object detection là phát hiện ra vùng ảnh chứa vật thể và gán nhãn phù hợp cho chúng. Tuy nhiên tiêu chuẩn về độ chính xác của Image Segmentation ở mức cao hơn so với Object Detection khi nó yêu cầu nhãn dự báo đúng tới từng pixel. Input của bài toán là một bức ảnh và output là một ma trận mask mà giá trị của từng pixel đã được gãn nhãn trên đó.

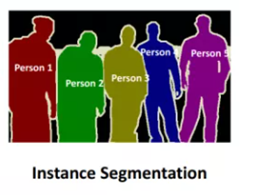
Bài toán phân đoạn hình ảnh có thể được chia làm 2 loại:

* Phân đoạn ngữ nghĩa (Sematic Segmentation) [39]: phân đoạn các vùng ảnh theo những nhãn khác nhau mà không phân biệt sự khác nhau giữa các đối tượng trong từng nhãn. Ví dụ trong bức ảnh xuất hiện 5 người, mức độ phân chia sẽ không xác định từng pixel thuộc về người nào.



Hình 10. Kết quả của bài toán phân đoạn ngữ nghĩa [39]

* Phân đoạn cá thể (Insane Segmentation) [39]: phân đoạn các vùng ảnh chi tiết đến từng đối tượng trong mỗi nhãn. Ví dụ: một ảnh có 5 người sẽ được phân chia segment chi tiết tới từng người 1, 2, … , 5.



Hình 11. Kết quả của bài toán Phân đoạn cá thể [39]

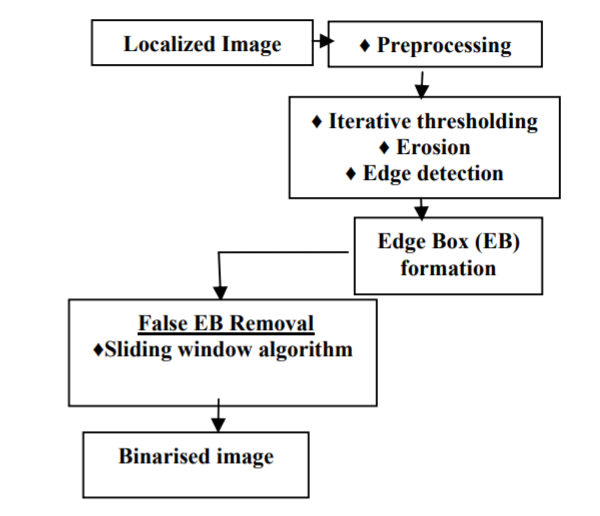
### **2.1.2. Các phương pháp phát hiện vùng kí tự**

Bài toán phát hiện và nhận dạng văn bản đã có những phương pháp tối ưu hoá và đưa ra kết quả rất tốt trong những năm gần đây. Trong phần này, các phương pháp truyền thống và hiện đại với việc sử dụng học sâu trong việc phát hiện các vùng kí tự sẽ được giới thiệu cùng những ưu điểm, nhược điểm của chúng.

#### **2.1.2.1. Phương pháp truyền thống**

Trình phát hiện văn bản truyền thống tập trung vào việc phát triển các đặc trưng cấp thấp được làm thủ công để phân biệt các thành phần văn bản và không phải văn bản trong hình ảnh, một trong những phương pháp phổ biễn là cửa sổ trượt (Sliding Windows - SW).

Việc nhận diện và phân loại các pixel trên ảnh là thuộc ký tự hay nền cùng với mã hoá nhị phân ảnh là bước tiền xử lý quan trọng góp phần tăng tỉ lệ chính xác của các thuật toán OCR. Các kỹ thuật hiện tại đang áp dụng bao gồm: Mã hoá nhị phân toàn cục và Mã hoá nhị phân cục bộ. Trong phương pháp mã hóa nhị phân toàn cục, ngưỡng toàn cục được sử dụng cho tất cả các pixel trong ảnh và không phù hợp với ảnh tài liệu phức tạp và đã xuống cấp. Phương thức này tỏ ra nhanh chóng và mạnh mẽ cho các đoạn văn bản nhỏ. Mặt khác, các phương pháp mã hóa nhị phân cục bộ thay đổi ngưỡng sao cho thích ứng với từng vùng trên hình ảnh. Phương pháp mã hóa nhị phân cục bộ được đề xuất để khắc phục các nhược điểm gặp phải của phương pháp toàn cục. Phương pháp mã hóa nhị phân cục bộ có thể được cải thiện bằng cách tính toán các ngưỡng cục bộ trong các cửa sổ hoặc khu vực riêng biệt. Trong hầu hết các phương pháp này, kích thước và hình dạng của cửa sổ là các tham số được xác định trước. Quá trình mã hóa nhị phân có thể thu lại kết quả kém khi ranh giới của cửa sổ vượt qua các ký tự và có thể làm phát sinh các ký tự bị hỏng và những khoảng trống (dấu cách), điều này có thể gây ra các lỗi không mong muốn trong quá trình OCR hình ảnh nhị phân.



Hình 12. Các bước của kỹ thuật nhận diện bằng cửa sổ trượt [40]

Đế áp dụng thuật toán cửa số trượt vào ảnh, cần các bước tiền xử lý:

- Tăng cường độ tương phản có điều kiện (Conditional Contrast Enhancement): Hình ảnh được chụp trong điều kiện ánh sáng kém có thể dẫn đến entropy thấp, điều này cần tăng độ tương phản của hình ảnh để xử lý tốt hơn. Do đó, entropy có thể được sử dụng như một chỉ báo nếu cần tăng độ tương phản của hình ảnh. Nếu entropy được tính toán của hình ảnh quá thấp, độ tương phản có thể được tăng lên, nếu không, các cạnh được phát hiện của các ký tự có thể không tạo thành các hình dạng khép kín.

- Làm mịn ảnh (Smothing): giảm thiểu ảnh hưởng của những nhiễu đối với quá trình phát hiện cạnh và thực hiện mở rộng thang màu xám cho hình ảnh được điều chỉnh để tăng thêm độ tương phản.

- Tạo ảnh xám (Grayscale extension): Có thể tính toán các giá trị thang độ xám tối đa và tối thiểu của hình ảnh được làm mịn. Nếu max\_S và min\_S gần nhau, ví dụ: max\_S - min\_S <80, hình ảnh thang độ xám sẽ đơn điệu với độ tương phản thấp. Phần mở rộng thang độ xám sẽ tăng độ tương phản thông qua khai thác toàn dải max\_S - min\_S = 255

- Áp dụng ngưỡng lặp lại (Iterative thresholding) để tạo ảnh nhị phân và xác định cạnh (Edge detection): Các cạnh sẽ được phát hiện từ hình ảnh được xử lý từ các bước trên. Các thuật toán phát hiện cạnh thông thường thuộc về bộ lọc thông cao, không phù hợp với các ảnh nền phức tạp. Ngưỡng tổng thể cố định thường sẽ hoạt động tốt đối với hình ảnh có nền đồng nhất, nhưng không hoạt động đối với nền có kết cấu và độ tương phản của các đối tượng thay đổi trong ảnh. Cũng cần duy trì tính liên tục của cạnh và tránh chồng chéo cạnh. Một thuật toán ngưỡng lặp lại và thuật toán xói mòn hình thái học được sử dụng ở đây để phát hiện các cạnh. Sau đó, xói mòn hình thái học được áp dụng để đảm bảo rằng các cạnh được phát hiện chỉ rộng một. Trong khi đó, điều này tránh được sự chồng chéo cạnh do tăng chiều rộng cạnh. Toán tử xói mòn sẽ loại bỏ tất cả các điểm biên khỏi một đối tượng và tất cả các điểm biên đều có chiều rộng là một pixel duy nhất. Nếu hình ảnh bị xói mòn bị trừ đi bởi hình ảnh gốc, kết quả sẽ là một cạnh rộng một pixel. Mỗi thành phần thu được, được gọi là một hộp cạnh (Edge Box - EB)

Các ký tự sẽ có kích thước xấp xỉ đồng nhất trong khu vực. Thuộc tính này được sử dụng để loại bỏ các EB không phải ký tự bằng cách phân tích nhị phân dựa trên phương pháp tiếp cận Cửa sổ trượt. Các bước liên quan đến thuật toán như sau:

a) Kích thước của các EB khác nhau được tính toán và lưu trữ trong mảng A [].

b) Ban đầu cửa sổ trên các phần tử mảng được giả định bằng cách bao phủ hai phần tử đầu tiên A [i] dưới dạng con trỏ trái (LP) và A [i + 1] dưới dạng con trỏ phải (RP).

c) Nếu sự khác biệt giữa LP và RP lớn hơn ngưỡng, chỉ xem xét RP và đặt nó là LP và trượt kích thước cửa sổ để che phần tử tiếp theo và kiểm tra sự khác biệt giữa hai phần tử đó.

d) So sánh lần lượt điểm đánh dấu cuối bên trái với các phần tử còn lại và nếu sự khác biệt < ngưỡng, mở rộng cửa sổ bằng cách đưa các phần tử đó vào cửa sổ và đóng băng cửa sổ khi điều kiện không thoả mãn. Cửa sổ này chứa các EB ký tự có kích thước thống nhất được xác định giữa LP và RP.

e) Sau đó, tính toán tối thiểu của các phần tử trong cửa sổ được đóng băng đầu tiên đó (MW1). Ở đây, giá trị tối thiểu được chọn để cho phép các ký tự có kích thước hỗn hợp.

f) Sau đó trượt cửa sổ để kiểm tra các phần tử còn lại và xác định các cửa sổ hợp lệ khác (Wi) và giá trị nhỏ nhất của các cửa sổ đó là MW1, MW2, … MWi

g) Tính toán tối thiểu MWi để cho kích thước hộp cạnh đồng nhất là ngưỡng Ts.

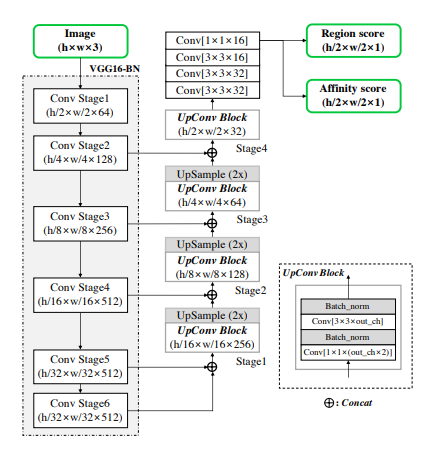
Bây giờ chia đôi hình ảnh với ngưỡng (Ts). Các EB có kích thước nhỏ hơn Ts bị loại bỏ. Pixel văn bản nền trước được hiển thị dưới dạng màu trắng và nền là pixel màu đen.

EB = 1 {Nếu kích thước (EB)> = Ts}

EB = 0 {Nếu kích thước (EB) <= Ts}

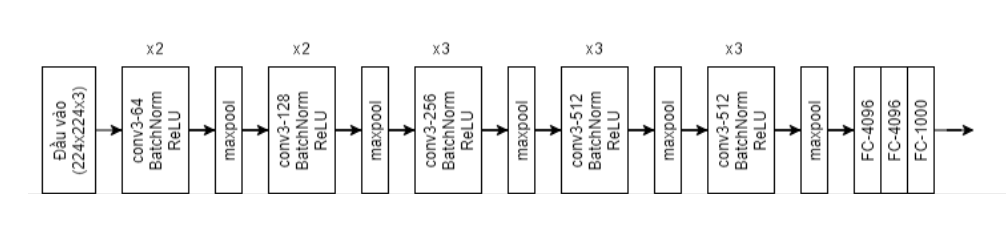
#### **2.1.2.2. Phương pháp học sâu với mô hình CRAFT (Character Region Awareness For Text)**

Mô hình CRAFT là một mô hình phát hiện văn bản dựa trên phân đoạn ngữ nghĩa được đề xuất bởi nhóm tác giả đến từ Clova AI Research, NAVER Corp, được thiết kế với một mạng nơ-ron phức hợp trả về điểm số khu vực (region score) và điểm mối quan hệ (affinity score). Điểm số khu vực được sử dụng để nhận diện các ký tự riêng lẻ trong hình ảnh và điểm số quan hệ được sử dụng để nhóm từng ký tự thành một từ. Mô hình này là một kiến trúc mạng phức hợp dựa trên VGG-16 với chuẩn hóa theo mẻ (batch normalization) được sử dụng làm phần khung. Mô hình sử dụng kết nối bỏ qua trong phần giải mã, tương tự như U-Net ở chỗ nó tổng hợp các đặc điểm ở mức thấp.



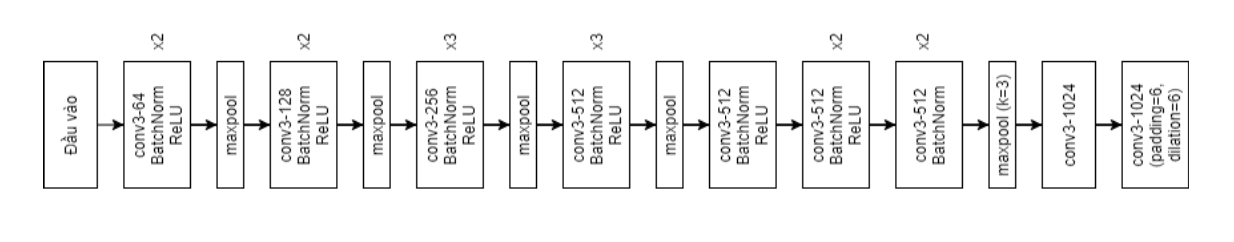
Hình 13. Kiến trúc mô hình CRAFT [41]

Từ hình ta thấy phần khung trích xuất đặc trưng của CRAFT là mạng VGG16 Batch Normalization, mạng mà thêm vào giữa mỗi lớp tích chập và lớp maxpooling một lớp chuẩn hóa theo mẻ (batch norm) được thể hiện trên hình 14:



Hình 14. Kiến trúc mạng VGG16BN [41]

VGG16 Batch Normalization áp dụng trong CRAFT có một số thay đổi ở điểm nó đã bỏ đi các lớp kết nối đầy đủ và xóa đi 1 lớp nhân chập ở nhóm cuối đồng thời thêm 1 lớp maxpool với kích thước bộ lọc là 3, tiếp đó là 1 lớp nhân chập conv3-1024 và cuối cùng là 1 lớp nhân chập conv3-1024 với vùng đệm bằng 6 và khoảng cách giữa các phần tử trong bộ lọc bằng 6 như trên hình 15.



Hình 15. Mạng VGG16BN áp dụng trong CRAFT [41]

Kết quả đầu ra của mạng khung (gọi là phần thu hẹp) là một tensor có kích thước h/32 x 2/32 x 512 với kích thước được giảm đi nhưng độ sâu được tăng lên. Tiếp theo đến phần mở rộng gồm các layer đối xứng tương ứng với các layer của phần thu hẹp. Quá trình Upsampling với 1 khối mạng tích chập và chuẩn hóa được áp dụng giúp cho kích thước layer tăng dần lên. Cuối cùng đầu ra của mô hình là ma trận điểm vùng và ma trận ái lực đều có kích thước h/2 x w/2. Mỗi điểm số vùng thể hiện xác suất pixel đã cho là trung tâm của ký tự và điểm số mối quan hệ biểu thị xác suất của pixel là trung tâm của khoảng cách giữa các ký tự liền kề. Sử dụng một số phương pháp hậu xử lý để nhận được kết quả cuối cùng là các hộp giới hạn bao quanh mỗi kí tự, từ hay dòng tùy theo thiết lập.

### 

### **2.1.3. Các phương pháp nhận dạng chuỗi kí tự**

Phần này giới thiệu các phương pháp đang được sử dụng nhiều trong nhận dạng chuỗi kí tự như CRNN dựa trên hàm mất mát CTC và TBRA dựa trên phương pháp Attention.

#### **2.1.3.1. Phương pháp CRNN**

Mô hình CRNN là một mô hình được đề xuất trong bài báo “*An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition*” bởi nhóm tác giả đến từ trường đại học và khoa học công nghệ Huazhong. CRNN là một kiến trúc mạng tích hợp các phần: trích xuất đặc trưng, mô hình trình tự và phiên mã thành vào một khung thống nhất. Theo tác giả, so với các hệ thống trước đây để nhận dạng văn bản cảnh, kiến ​​trúc CRNN sở hữu bốn đặc tính đặc biệt:

(1) Là một thuật toán được huấn luyện liền mạch từ đầu đến cuối, trái ngược với hầu hết các thuật toán hiện có mà các thành phần được huấn luyện và điều chỉnh riêng biệt.

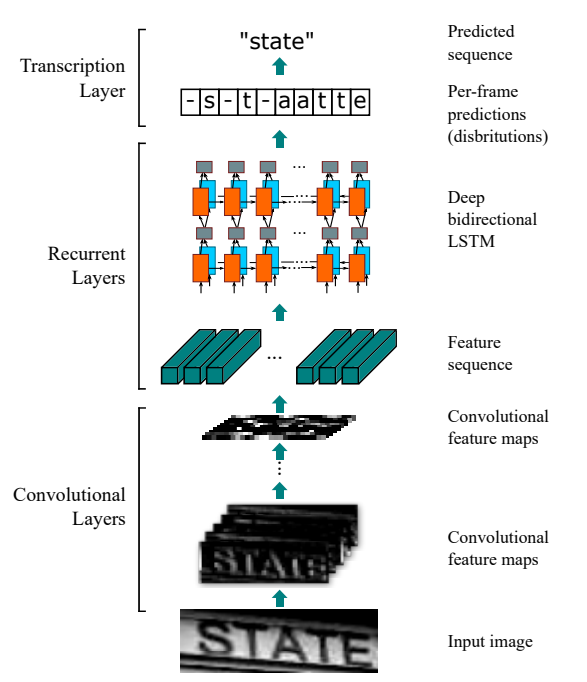
(2) Nó xử lý các chuỗi với độ dài tùy ý, không liên quan đến việc phân đoạn kí tự hoặc chuẩn hóa theo tỷ lệ ngang.

(3) Nó không bị giới hạn trong bất kỳ từ điền nào cho trước và đạt được hiệu suất tốt đối với những văn bản cấu thành từ từ vựng (lexincon-based) hay từ các từ viết tắt, từ đặc biệt (lexicon-free).

(4) Nó tạo ra một mô hình hiệu quả nhưng nhỏ hơn nhiều, dễ áp dụng hơn cho các tình huống ứng dụng trong thế giới thực.

Các thử nghiệm trên các điểm chuẩn tiêu chuẩn, bao gồm bộ dữ liệu IIIT-5K [21], các dữ liệu văn bản đường phố và ICDAR [22], chứng minh tính ưu việt của thuật toán được đề xuất so với các kỹ thuật trước đây.

Kiến trúc mạng của CRNN, như trong Hình 16, từ dưới lên bao gồm ba thành phần: các lớp tích chập (Convolutional Layers), các lớp hồi quy (Recurrent Layers) và một lớp phiên mã (Transcription Layer). Ở cuối CRNN, các lớp tích chập tự động trích xuất chuỗi đặc điểm từ mỗi hình ảnh đầu vào. Tiếp theo, một mạng hồi quy được xây dựng để đưa ra dự đoán cho từng khung của chuỗi đặc trưng. Lớp phiên mã ở đầu CRNN được sử dụng để dịch các dự đoán trên mỗi khung hình bởi các lớp hồi quy thành một chuỗi nhãn. Mặc dù CRNN bao gồm các loại kiến trúc mạng khác nhau (ví dụ: CNN và RNN), nó có thể được huấn luyện cùng với một hàm mất mát.

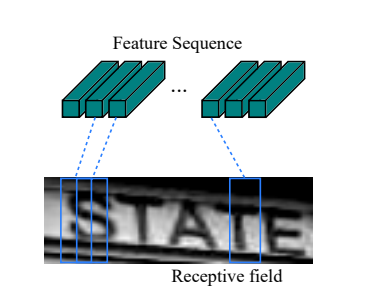


Hình 16. Kiến trúc CRNN [42]

##### **a. Trích xuất đặc trưng**

Trong mô hình CRNN, thành phần của các lớp tích chập được xây dựng bằng cách lấy các lớp tích chập và pooling từ mô hình CNN tiêu chuẩn (các lớp fully-connected sẽ bị loại bỏ). Những thành phần này sẽ được sử dụng để trích xuất một biểu diễn đặc trưng từ một hình ảnh đầu vào. Trước khi được đưa vào mạng, tất cả các hình ảnh cần được thu nhỏ đến cùng một kích thước. Sau đó, một chuỗi các vectơ đặc trưng được trích xuất từ ​​các bản đồ đặc trưng được tạo ra bởi thành phần của các lớp tích chập, là đầu vào cho các lớp hồi quy. Cụ thể, mỗi vectơ đặc trưng của chuỗi đối tượng được tạo từ trái sang phải kết nối với nhau theo các cột. Điều này có nghĩa là vectơ đặc trưng thứ i là sự ghép nối của các cột thứ i của cả bản đồ. Chiều rộng của mỗi cột được cố định thành một pixel. Khi các lớp của hàm tích chập, max-pooling và chức năng kích hoạt theo từng phần tử hoạt động trên các vùng cục bộ, chúng là bất biến. Do đó, mỗi cột của bản đồ đối tượng tương ứng với một vùng hình chữ nhật của hình ảnh gốc (được gọi là trường tiếp nhận) và các vùng hình chữ nhật đó có cùng thứ tự với các cột tương ứng của chúng trên bản đồ đối tượng từ trái sang phải Do đó, mỗi cột của bản đồ đặc trưng tương ứng với một vùng hình chữ nhật của hình ảnh gốc (được gọi là trường tiếp nhận) và các vùng hình chữ nhật đó có cùng thứ tự với các cột tương ứng của chúng trên bản đồ đối tượng từ trái sang phải. Như được minh họa trong hình 17, mỗi vectơ trong chuỗi đặc trưng được liên kết với một trường tiếp nhận và có thể được coi là bộ mô tả hình ảnh cho khu vực đó.

Do tính mạnh mẽ, phong phú và có thể huấn luyện nên các đặc trưng tích chập đã được áp dụng rộng rãi cho các bài toán nhận dạng hình ảnh khác nhau [19, 20]. Một số phương pháp tiếp cận trước đây đã sử dụng CNN để tìm hiểu cách biểu diễn mạnh mẽ cho các đối tượng dạng chuỗi như văn bản cảnh [21]. Tuy nhiên, các cách tiếp cận này thường trích xuất đại diện tổng thể của toàn bộ hình ảnh bởi CNN, sau đó các đặc điểm sâu cục bộ được thu thập để nhận ra từng thành phần của một đối tượng tương tự. Vì CNN yêu cầu các hình ảnh đầu vào phải được chia tỷ lệ thành một kích thước cố định để đáp ứng với kích thước đầu vào cố định của nó, nó không thích hợp cho các đối tượng dạng chuỗi do sự thay đổi độ dài lớn của chúng. Trong CRNN, các đặc trưng sâu được truyền vào các biểu diễn tuần tự để chúng trở nên bất biến với sự thay đổi độ dài của các đối tượng giống chuỗi.



Hình 17. Trường tiếp nhận. Mỗi vectơ trong chuỗi đặc trưng được trích xuất được liên kết với một trường tiếp nhận trên hình ảnh đầu vào và có thể được coi là vectơ đặc trưng của trường đó [42].

##### **b. Tạo chuỗi trình tự**

Như đã được đề cập tại phần giới thiệu, ở bước này áp dụng một mạng neuron hồi quy (Recurrent Neural Network - RNN) để tìm ra phân phối xác suất đầu ra từ các chuỗi đặc trưng tại bước trước. Trong đó mỗi đơn vị được sử dụng trong RNN là LSTM nên ta cũng có thể gọi là mạng LSTM.

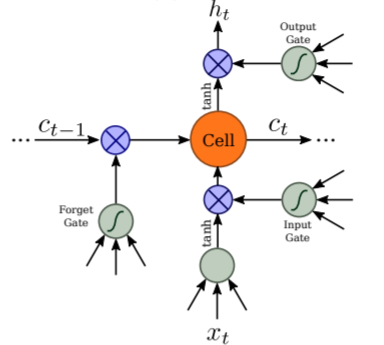
\* Mạng LSTM

LSTM là một phiên bản mở rộng của mạng RNN, được đề xuất vào năm 1997 bởi Sepp Hochreiter và Jürgen Schmidhuber [11]. LSTM được thiết kế để giải quyết các bài toán về phụ thuộc xa (long-term dependencies) trong mạng RNN do bị ảnh hưởng bởi vấn đề Gradient biến mất.

\* Gradient biến mất: Khi nhiều layer sử dụng các hàm kích hoạt được thêm vào mạng neuron, các gradient của hàm mất mát sẽ tiến về 0, làm cho mạng khó đào tạo. [43]

Một số hàm kích hoạt nhất định, chẳng hạn như hàm sigmoid, chuyển đổi một không gian đầu vào lớn thành một không gian đầu vào nhỏ có giá trị từ 0 đến 1. Do đó, một sự thay đổi lớn trong đầu vào của hàm sigmoid sẽ gây ra một thay đổi nhỏ trong đầu ra. Do đó, đạo hàm trở nên nhỏ. Đối với mạng nhỏ chỉ có một vài layer sử dụng các kích hoạt này, đây không phải là vấn đề lớn. Tuy nhiên, khi sử dụng nhiều layer hơn, nó có thể gây ra gradient quá nhỏ để quá trình đào tạo hoạt động hiệu quả.

Gradients của mạng neuron được tính bằng cách sử dụng lan truyền ngược (backpropagation). Nói một cách đơn giản, backpropagation tìm các dẫn xuất của mạng bằng cách di chuyển từng layer từ lớp cuối cùng đến lớp ban đầu. Theo quy tắc chuỗi, các dẫn xuất của mỗi lớp được nhân xuống mạng để tính các dẫn xuất của các layer ban đầu. Tuy nhiên, khi n layer ẩn sử dụng một kích hoạt như hàm sigmoid, n đạo hàm nhỏ sẽ được nhân với nhau. Do đó, gradient giảm theo cấp số nhân khi chúng ta truyền tới các lớp ban đầu. Một gradient nhỏ có nghĩa là trọng số và độ lệch của các lớp ban đầu sẽ không được cập nhật một cách hiệu quả. Vì các lớp ban đầu này thường rất quan trọng để nhận ra các yếu tố cốt lõi của dữ liệu đầu vào, nó có thể dẫn đến sự thiếu chính xác chung của toàn mạng. Có thể hiểu một cách đơn giản là mạng RNN cơ bản trong thực tế không có khả năng ghi nhớ thông tin từ các bước có khoảng cách xa và do đó những phần tử đầu tiên trong chuỗi đầu vào không có nhiều ảnh hưởng đến các kết quả tính toán dự đoán phần tử cho chuỗi đầu ra trong các bước sau.



Hình 18. : Cấu trúc bên trong tế bào LSTM

Mạng LSTM có thể bao gồm nhiều tế bào LSTM (LSTM memory cell) liên kết với nhau và kiến trúc cụ thể của mỗi tế bào được biểu diễn như trong Hình 18. Bao gồm một ô nhớ và ba cồng: cổng vào, cổng ra và cổng quên.

Ở state thứ t của mô hình LSTM:

- Output:  ta gọi c là cell state, h là hidden state.

- Input: trong đó  là input ở state thứ t của model; là output của layer trước.



Hình 19. Mô hình hoạt động của LSTM [44]

​ tương ứng với **f**orget gate, **i**nput gate và **o**utput gate.

- Forget gate:

- Input gate:

- Output gate:

Với: 0 <, , < 1; , ,  là các hệ số bias.

Ta có:

, **forget gate** quyết định xem cần lấy bao nhiêu từ cell state trước và **input gate** sẽ quyết định lấy bao nhiêu từ input của state và hidden layer của layer trước.

, **output gate** quyết định xem cần lấy bao nhiêu từ cell state để trở thành output của hidden state. Ngoài ra  cũng được dùng để tính ra output  ​ cho state t.

Các kí hiệu được sử dụng trong các công thức trên:

+ là véc tơ đầu vào tại mỗi bước thời gian *t.*

*+* , , , , , , , là các ma trận trong mỗi tế bào LSTM.

+ , , , là các véc tơ độ lệch.

+ , , chứa các giá trị kích hoạt lần lượt cho các cổng quên, cổng vào, cổng ra.

+ , lần lượt là các véc tơ đại diện cho trạng thái nội bộ tế bào và giá trị ứng viên.

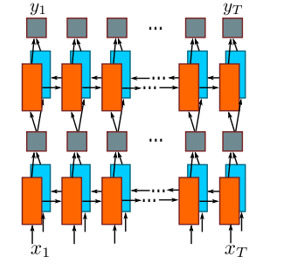
\* Áp dụng trong mô hình

Mạng neuron hồi quy hai chiều sâu (BiLSTM) được xây dựng để nhận đầu ra của các lớp tích chập. Các lớp hồi quy dự đoán phân phối nhãn cho mỗi khung trong chuỗi đặc trưng . Mạng này có những ưu điểm như sau :

- Thứ nhất, RNN có thế hiện khả năng mạnh mẽ trong việc nắm bắt thông tin theo ngữ cảnh. Sử dụng các dấu hiệu theo ngữ cảnh để nhận dạng chuỗi dựa trên hình ảnh ổn định và hữu ích hơn việc xử lý từng ký hiệu một cách độc lập. Bên cạnh đó, một số ký tự không rõ ràng sẽ dễ phân biệt hơn khi quan sát ngữ cảnh của chúng, ví dụ: dễ dàng nhận ra "il" bằng cách đối chiếu chiều cao của ký tự hơn là nhận ra từng ký tự một cách riêng biệt.

- Thứ hai, RNN có thể áp dụng thuật toán lan truyền ngược để có thể sửa những sai lệch từ đầu vào của nó, có thể thông qua các lớp tích chập, cho phép chúng ta cùng huấn luyện các lớp hồi quy và các lớp tích chập trong một mạng thống nhất.

- Thứ ba, RNN có thể hoạt động trên các chuỗi có độ dài tùy ý.



Hình 20. Cấu trúc của LSTM hai chiều sâu.

Kết hợp LSTM xuôi (trái sang phải) và ngược (phải sang trái). [43]

Một đơn vị RNN truyền thống có một lớp ẩn tự kết nối giữa các lớp đầu vào và đầu ra của nó. Mỗi khi nó nhận được một khung trong chuỗi, nó sẽ cập nhật trạng thái ẩn của nó bằng một hàm phi tuyến tính lấy cả đầu vào hiện tại và trạng thái trong quá khứ làm đầu vào của nó . Sau đó, dự đoán được thực hiện dựa trên . Bằng cách này, các ngữ cảnh trong quá khứ được ghi lại và sử dụng để dự đoán. Tuy nhiên, như đã được đề cập, đơn vị RNN truyền thống gặp phải vấn đề Gradient biến mất, giới hạn phạm vi bối cảnh mà nó có thể lưu trữ và thêm gánh nặng cho quá trình huấn luyện nên tác giả đã sử dụng LSTM, cho phép nó nắm bắt các phụ thuộc xa, thường xảy ra trong các chuỗi dựa trên hình ảnh. Từ định hướng là việc sử dụng LSTM, tác giải nhận thấy trong trình tự dựa trên hình ảnh, bối cảnh từ cả hai hướng đều hữu ích và bổ sung cho nhau. Do đó, tác giả kết hợp hai LSTM, một chiều xuôi một chiều ngược, thành một LSTM hai chiều. Hơn nữa, nhiều LSTM hai chiều có thể được xếp chồng lên nhau, dẫn đến một LSTM hai chiều sâu như minh họa trong Hình 20. Cấu trúc sâu cho phép mức độ trừu tượng cao hơn cấu trúc nông và đã đạt được những cải tiến hiệu suất đáng kể trong nhiệm vụ nhận dạng giọng nói [3].

##### **c. Sự phiên mã**

Phiên mã là quá trình chuyển đổi các dự đoán trên mỗi khung hình do RNN trả về thành một chuỗi trình tự. Về mặt toán học, phiên mã là để tìm chuỗi trình tự có xác suất cao nhất được điều chỉnh dựa trên các dự đoán trên mỗi khung hình. Trong thực tế, tồn tại hai hình thức phiên mã, đó là phiên mã không có từ vựng và dựa trên từ vựng. Ở chế độ không có từ vựng, các dự đoán được thực hiện mà không có bất kỳ từ vựng nào. Trong chế độ dựa trên từ vựng, các dự đoán được thực hiện bằng cách chọn chuỗi nhãn có xác suất cao nhất.

\* Xác suất của chuỗi trình tự

Tác giả áp dụng xác suất có điều kiện được xác định trong Phân loại theo thời gian kết nối (Connectionist temporal classification - CTC) do Graves và các cộng sự đề xuất [9]. Xác suất có điều kiện của nhãn trình tự L được xác định dựa trên các dự đoán trên mỗi khung hình , và nó bỏ qua vị trí của mỗi nhãn trong L. Do đó, khi chúng ta sử dụng hàm negative log-likelihood của xác suất này làm hàm mất mát để huấn luyện mạng, chúng ta chỉ cần hình ảnh và chuỗi nhãn tương ứng của chúng, tránh tốn công gắn nhãn vị trí của các ký tự riêng lẻ.

Giả sử đầu ra của mạng là một chuỗi trong đó T là độ dài của chuỗi, trong đó với là xác suất lớn nhất của một phân phối xác suất trên tập nhãn kí tự và là xác suất của kí tự thứ k tại thời điểm t, trong đó L chứa toàn bộ kí tự nhãn và kí tự ‘blank’ là kí tự trống được thêm vào. Gọi là một chuỗi thuộc tập , các chuỗi có độ dài T với mỗi kí tự thuộc . Xác suất có điều kiện của mỗi chuỗi , được gọi là chuỗi căn chỉnh (alignment), với chuỗi đặc trưng đầu vào *x* là tích của các biến cố trong (giả định các biến cố độc lập nhau):

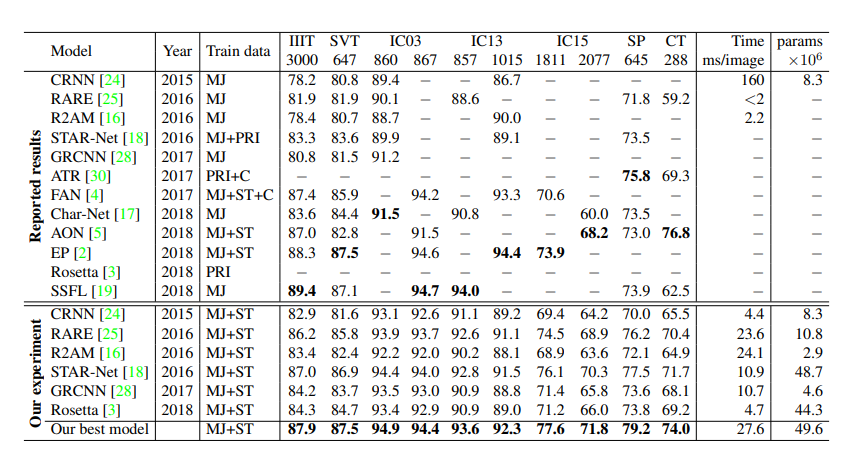
Ví dụ các chuỗi căn chỉnh từ nhãn “hello” là “heee---llo”, “-hell--loo”… Tiếp theo, tác giả định nghĩa hàm ánh xạ B : , trong đó là tập các chuỗi trình tự đầu ra có độ dài nhỏ hơn hoặc bằng T sau khi đã loại bỏ các kí tự lặp lại và kí tự “blank”, ví dụ . Xác suất có điều kiện của kết quả *l* (labelling) với đầu vào *x* được định nghĩa là tổng xác suất có điều kiện của các chuỗi trình tự với đầu vào *x* qua hàm B được kết quả *l*

(1)

Kết quả đầu ra thỏa mãn công thức:

#### **2.1.3.2. Phương pháp TRBA**

Năm 2019 nhóm tác giả đến từ Clova AI Research đã công bố bài báo “What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis” [12] đề cập về sự không nhất quán của các tập dữ liệu huấn luyện và đánh giá cho những mô hình nhận dạng văn bản cảnh (scene text recognition) được đưa ra những năm gần đây mà có kết quả nhận dạng kém. Và họ đã giới thiệu một khuôn mẫu thống nhất bao gồm bốn bước áp dụng cho các mô hình nhận dạng văn bản cảnh. Họ cũng đưa ra đánh giá về nhiều mô hình như CRNN [5], RARE [10], R2AM [8], STAR-Net [14], GRCNN [15], Rosetta [16],…và mô hình TRBA (TPS-ResNet-BiLSTM-Attn) do họ đề xuất, mô hình nhận dạng văn bản cảnh (STR) có độ chính xác cao nhất trên các tập dữ liệu được kiểm nghiệm trên hình 21.



Hình 21. So sánh hiệu quả của các mô hình STR hiện có và thực nghiệm trên các mô hình đó của nhóm tác giả [12]

Các bước mà nhóm tác giả đề cập bao gồm:

- Chuyển đổi: chuẩn hóa hình ảnh chuỗi kí tự đầu vào bằng cách sử dụng mạng biến áp không gian (STN) để nắn những chuỗi kí tự bị cong, nghiêng trên ảnh.

- Trích xuất đặc trưng: sử dụng mạng nơ ron tích chập để ánh xạ hình ảnh đầu vào thành những đặc trưng mà liên quan đến nhận dạng kí tự, đồng thời loại bỏ đồng thời loại bỏ các đặc trưng không liên quan như phông chữ, màu sắc, kích thước và nền.

- Tạo trình tự: sử dụng mạng nơ ron hồi quy (RNN) để nắm bắt thông tin ngữ cảnh trong một chuỗi ký tự cho bước tiếp theo để dự đoán từng ký tự một cách tốt hơn, thay vì chỉ sử dụng chuỗi đặc trưng từ bước trước đó.

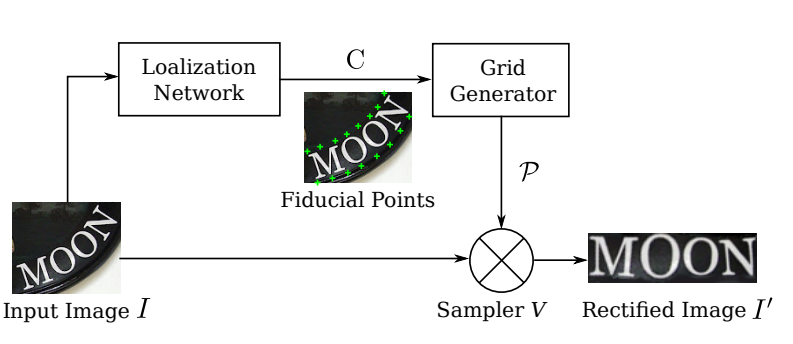
- Dự đoán: dự đoán chuỗi ký tự đầu ra từ các đặc điểm đã xác định của hình ảnh.

##### **a. Bước chuyển đổi**

Giai đoạn này biến hình ảnh đầu vào thành hình ảnh được chuẩn hóa . Các hình ảnh chuỗi kí tự thường có nhiều hình dạng cong, nghiêng khác nhau. Nếu những hình ảnh đầu vào như vậy không được xử lý thì giai đoạn trích xuất trích xuất đặc trưng tiếp theo cần phải học cách biểu diễn của các chuỗi kí tự với nhiều hình dạng khác nhau. Để giảm bớt gánh nặng này, phép biến đổi spline bản mỏng (TPS), đã được áp dụng để chuẩn hóa hình ảnh đầu vào.

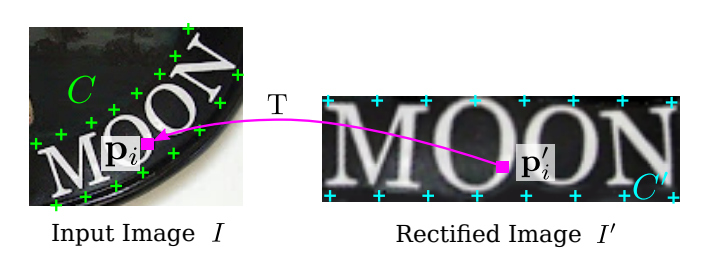
\* Mạng STN với phép biến đổi TPS

STN [7], được đề xuất bởi Max Jaderberg và các cộng sự năm 2015, là một kiến trúc mạng bao gồm một chuỗi các quy trình: tìm đường biên văn bản, liên kết vị trí của các pixel trong đường biên với vị trí của hình ảnh chuẩn hóa và tạo hình ảnh chuẩn hóa bằng cách sử dụng các giá trị của pixel và thông tin liên kết. Các quy trình như vậy được gọi tương ứng là mạng cục bộ hóa, bộ tạo lưới và bộ lấy mẫu hình ảnh và được thể hiện trên hình…. Về mặt khái niệm, TPS sử dụng phép nội suy đường thẳng giữa một tập hợp các điểm cơ sở (fiducial points) đại diện cho đường viền xung quanh của chuỗi kí tự trong một hình ảnh, số lượng các điểm cơ sở là F và nó là một hằng số.



Hình 22. Kiến trúc của STN. Mạng cục bộ hóa định vị các điểm cơ sở C, bộ tạo lưới tạo ra lưới lấy mẫu P, bộ lấy mẫu V tạo hình ảnh chuẩn hóa I' từ hình ảnh đầu vào I và P [10]

Mạng cục bộ hóa là một mạng CNN bao gồm các lớp tích chập, các lớp tổng hợp và lớp kết nối đầy đủ. Lớp đầu ra của mạng, lớp kết nối đầy đủ cuối cùng, được thiết lập với số lượng đầu ra là 2F sau đó được đưa ra hàm kích hoạt *tanh* để đưa các giá trị đầu ra nằm trong [-1, 1], F là số điểm cơ cở trên hình ảnh đầu vào và các điểm cơ sở được biểu thị bởi ma trận trong đó cột thứ f chứa tọa độ của điểm cơ sở thứ f. Mạng cục bộ hóa định vị các điểm cơ sở trên toàn ảnh, nó được mong đợi có thể nắm bắt hình dạng chuỗi kí tự tổng thể của hình ảnh đầu vào và khoanh vùng các điểm cở sở cho phù hợp.



Hình 23. Tập điểm cơ sở và bộ chuyển đổi TPS [10]

Bộ tạo lưới tính toán những tham số chuyển đổi TPS và tạo ra lưới lấy mẫu. Đầu tiên định nghĩa trước một tập các điểm cơ sở của hình ảnh chuẩn hóa biểu thị bởi ma trận . Như được minh họa trên hình … bộ tạo lưới cung cấp một hàm ánh xạ các vùng được xác định ở mạng cục bộ hóa và hình ảnh chuẩn hóa. Hàm ánh xạ được tham số hóa bởi ma trận T

(1)

trong đó là một ma trận được xác định bởi C, do đó nó cũng là hằng

(2)

trong đó phần tử hàng i cột j của R có giá trị , là khoảng cách euclidean giữa và . Lưới pixels trên ảnh chuẩn hóa được biểu thị bằng , trong đó là tọa độ của pixel thứ i, N là số lượng pixel. Trên hình 3, với mỗi pixel trong , thực hiện tìm điểm trong *I* bằng việc áp dụng phép chuyển đổi:

(3)

(4)

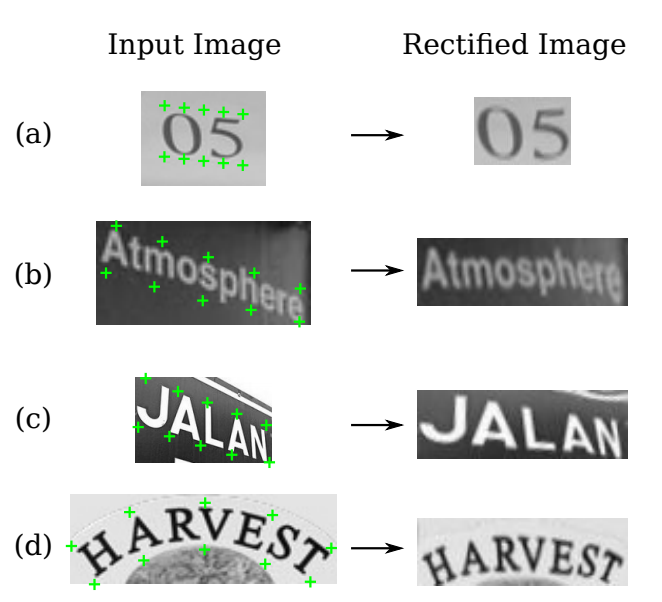
(5)

trong đó là khoảng cách euclidean giữa và . Thực hiện lặp lại công thức (3) trên toàn bộ các điểm của , tạo ra lưới trên *I.*

Cuối cùng, trong bộ lấy mẫu, những giá trị pixel của được nội suy song tuyến từ các điểm ảnh gần trên hình ảnh đầu vào. Thực hiện tính toán tất cả các giá trị điểm ảnh, nhận được hình ảnh đầu ra :

trong đó V đại diện cho bộ lấy mẫu song tuyến.

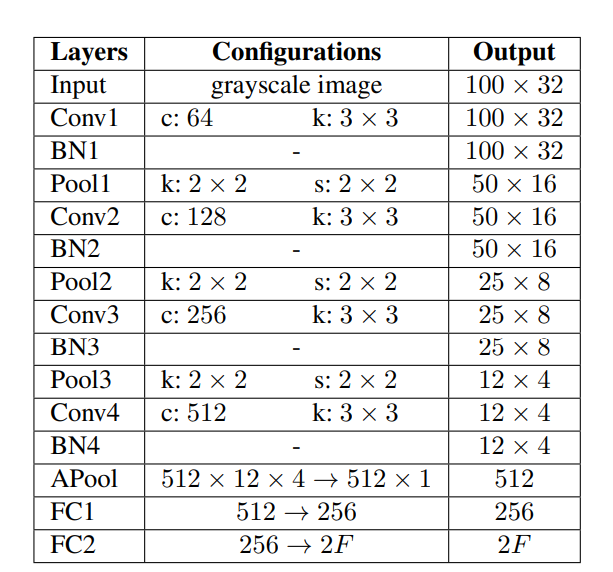
Tính linh hoạt của phép chuyển đổi TPS cho phép biến đổi các hình ảnh chuỗi kí tự không đúng chuẩn thành các hình ảnh chỉnh sửa có chứa chuỗi kí tự đúng chuẩn. Trong hình 24 là một số loại chuỗi kí tự không đúng chuẩn phổ biến, bao gồm a) chuỗi kí tự có hộp giới hạn không cắt sát, do trình phát hiện văn bản chưa tốt; b) chuỗi kí tự nhiều hướng, gây ra bởi các camera chụp không nằm ngang; c) văn bản phối cảnh, do góc máy ảnh xem bên; d) Chữ cong, một trường hợp thường thấy. STN có thể chỉnh sửa các hình ảnh chứa các loại chuỗi kí tự bất thường này, làm cho chúng dễ đọc hơn đối với trình nhận dạng sau.



Hình 24. STN căn chỉnh một số hình ảnh chuỗi kí tự thường gặp [10]

\* Áp dụng trong mô hình

STN yêu cầu mạng cục bộ tính toán các điểm cơ sở của hình ảnh đầu vào. Nhóm tác giả thiết kế mạng cục bộ bằng cách tuân theo hầu hết các thành phần của các nghiên cứu trước đó [7] và thêm các lớp chuẩn hóa theo mẻ và tổng hợp trung bình thích ứng (adaptive average pooling) để ổn định việc huấn luyện mạng. Hình 25 thể hiện kiến trúc mạng cục bộ bao gồm 4 lớp tích chập, mỗi lớp theo sau là lớp chuẩn hóa theo mẻ và lớp tổng hợp tối đa 2 x 2. Kích thước bộ lọc, kích thước phần đệm và bước nhảy tương ứng là 3, 1, 1 cho tất cả các lớp tích chập. Theo sau lớp chập cuối cùng là lớp gộp trung bình thích ứng. Sau đó, hai lớp được kết nối đầy đủ như sau: 512 đến 256 và 256 đến 2F. Đầu ra cuối cùng là véc tơ 2F chiều tương ứng với giá trị của tọa độ x, y của F điểm cơ sở trên hình ảnh đầu vào. Các hàm kích hoạt cho tất cả các lớp là ReLU



Hình 25. Kiến trúc của mạng cục bộ trong STN [12]

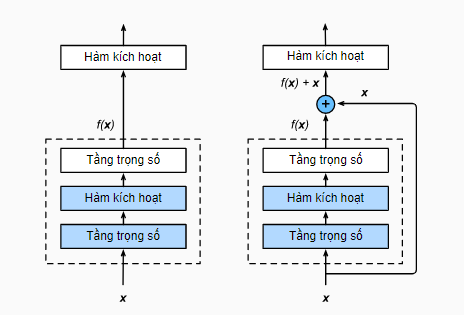
##### **b. Bước trích xuất đặc trưng**

Trong giai đoạn này, một mạng nơ ron tích chập sẽ trích xuất từ hình ảnh đầu vào và kết quả là một bản đồ đặc trưng (I là số lượng cột của bản đồ đặc trưng). Mỗi cột trong bản đồ đặc trưng kết quả được thực hiện bằng một trình trích xuất đặc trưng có một trường tiếp nhận có thể phân biệt được tương ứng dọc theo đường ngang của hình ảnh đầu vào. Các đặc trưng này được sử dụng để ước tính ký tự trên mỗi trường tiếp nhận. Nhóm tác giả nghiên cứu ba kiến ​​trúc VGG [4], RCNN [8], và ResNet [13]. VGG ở dạng ban đầu bao gồm nhiều lớp chập, tiếp theo là một vài lớp được kết nối đầy đủ. RCNN là một biến thể của CNN có thể được áp dụng đệ quy để điều chỉnh các trường tiếp nhận của nó tùy thuộc vào hình dạng ký tự. ResNet là CNN với các kết nối dư giúp giảm bớt việc huấn luyện các mạng CNN sâu hơn.

Trong phạm vi đồ án này sử dụng mạng ResNet tương ứng phần R trong mô hình TRBA.

\* Mạng ResNet

ResNet (Residual Network) [13] được giới thiệu vào năm 2015 và thậm chí đã giành được vị trí thứ 1 trong cuộc thi ILSVRC 2015 với tỉ lệ lỗi top 5 chỉ 3.57%. Không những thế nó còn đứng vị trí đầu tiên trong cuộc thi ILSVRC and COCO 2015 với ImageNet Detection, ImageNet localization, Coco detection và Coco segmentation. Hiện tại thì có rất nhiều biến thể của kiến trúc ResNet với số lớp khác nhau như ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152...Với tên là ResNet theo sau là một số chỉ kiến trúc ResNet với số lớp nhất định. Mạng ResNet được thiết kế để làm việc với hàng trăm hoặc hàng nghìn lớp chập. Một vấn đề xảy ra khi xây dựng mạng CNN với nhiều lớp chập sẽ xảy ra hiện tượng Gradient biến mất dẫn tới quá trình học tập không tốt. Cho nên giải pháp mà ResNet đưa ra là sử dụng kết nối "tắt" đồng nhất để xuyên qua một hay nhiều lớp. Một khối như vậy được gọi là một Residual Block, như trong hình sau:

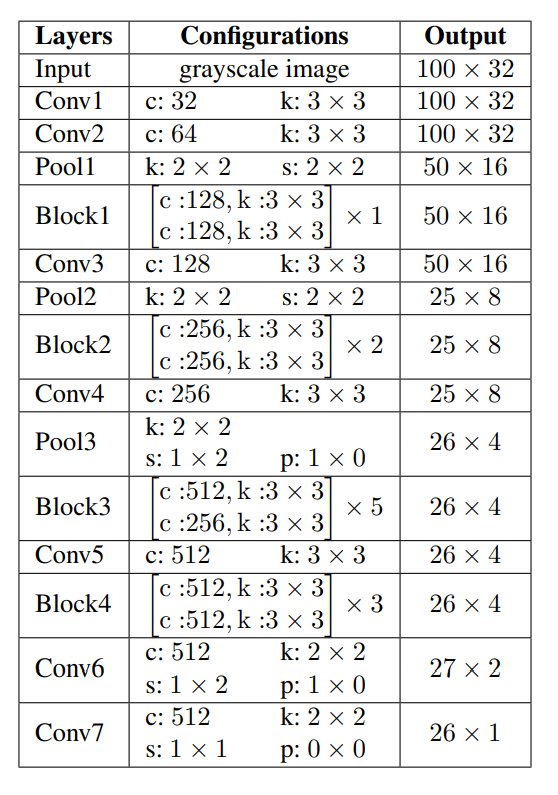


Hình 26. Sự khác biệt giữa khối thông thường (trái) và khối phần dư (phải) [23]

Từ mạng nơ ron trên hình 26, ký hiệu đầu vào là . Giả sử ánh xạ lý tưởng muốn học được là  và được dùng làm đầu vào của hàm kích hoạt. Phần nằm trong viền nét đứt bên trái phải khớp trực tiếp với ánh xạ . Điều này có thể không đơn giản nếu ta không cần khối đó và muốn giữ lại đầu vào x. Khi đó, phần nằm trong viền nét đứt bên phải chỉ cần tham số hoá độ lệch khỏi giá trị x, bởi vì ta đã trả về . Trên thực tế, ánh xạ phần dư thường dễ tối ưu hơn, vì chỉ cần đặt . Nửa bên phải mô tả khối phần dư cơ bản của ResNet.

\* Áp dụng trong mô hình

Nhóm tác giả triển khai mô đun ResNet với 29 lớp có thể huấn luyện. Trên hình 26 mô tả chi tiết về mạng được sử dụng, trong đó c là số kênh và k là kích thước bộ lọc và s là bước nhảy. Mỗi lớp Conv gồm các lớp theo thứ tự: 1 lớp tích chập, 1 lớp BatchNorm và 1 lớp ReLU. Mỗi lớp Block là các khối phần dư gồm các lớp theo thứ tự: 1 lớp tích chập , 1 lớp BatchNorm, 1 lớp relu, 1 lớp tích chập và 1 lớp BatchNorm rồi sau đó đầu ra của chuỗi lớp này được kết hợp với đầu vào ở lớp trước lớp Block rồi qua 1 lớp ReLU. Với đầu vào là ảnh xám có kích thước 100 x 32 tương ứng chiều rộng là 100 và chiều cao là 32 pixel thì đầu ra có kích thước là 26 x 1.



Hình 27. Mạng ResNet được sử dụng trong TBRA [12]

##### **c. Bước tạo trình tự**

Giai đoạn này giúp nắm bắt thông tin ngữ cảnh trong một chuỗi ký tự cho giai đoạn tiếp theo từ chuỗi đặc trưng V, được tạo thành từ bản đồ đặc trưng ở giai đoạn trích xuất đặc trưng, để dự đoán từng ký tự một cách tốt hơn thay vì làm điều đó một cách độc lập.

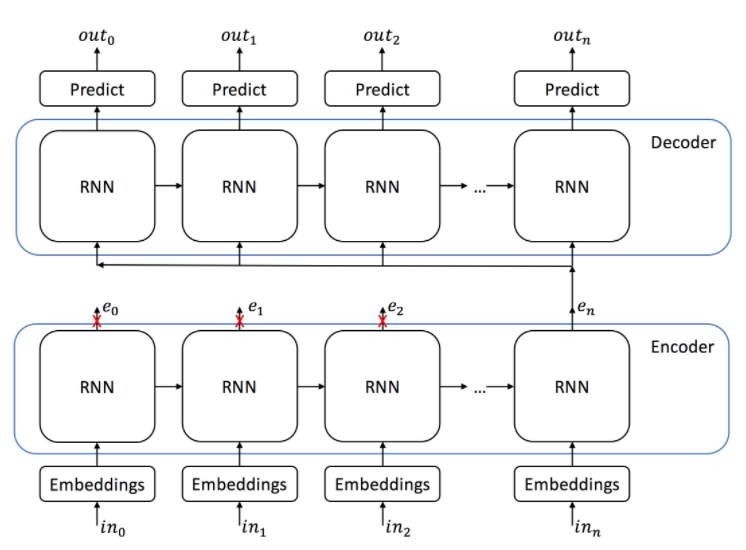
\* Áp dụng trong mô hình

Sử dụng 2 mạng LSTM hai chiều (BiLSTM), cũng được sử dụng trong CRNN [5] đã được đề cập. Mạng BiLSTM thứ *l* xác định hai trạng thái ẩn , được tính toán qua chuỗi trình tự và đảo ngược của nó. Để thu được đầu ra là phân phối xác suất của các các véc tơ đặc trưng tại bước trước.

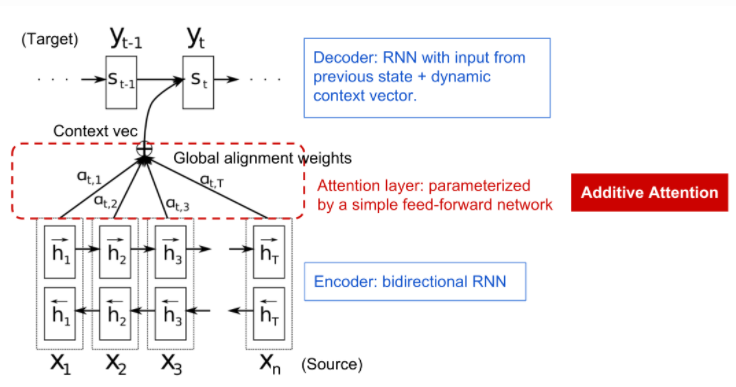
##### **d. Bước dự đoán kết quả**

Giai đoạn này kết hợp với giai đoạn tạo trình tự để tính toán kết quả đầu ra của chuỗi kí tự từ những đặc trưng của hình ảnh đã được xác định trước đó. Tại bước này tác giả triển khai bộ giải mã dựa trên cơ chế chú ý (attention-based decoder). Bộ giải mã nhằm tìm phân phối xác suất của các đặc trưng học tập (learning features) ở bước mã hóa, được tính toán qua mạng RNN tại bước tạo trình tự, từ đó xác định đâu là nhãn đầu ra. Mô hình mã hóa – giải mã trên thuộc lớp seq2seq tức ánh xạ chuỗi thành chuỗi.

Cơ chế chú ý (attention) được giới thiệu trong bài báo “Neural machine translation by jointly learning to align and translate” bởi Dzmitry Bahdanau và các cộng sự năm 2015 [17], nó được áp dụng đầu tiên trong mô hình dịch máy (Neural Machine Translation), theo tác giả các kí tự ở đầu vào sẽ có mối liên hệ lớn hơn đối với kí tự ở đầu ra cùng vị trí. Do đó attention hiểu một cách đơn giản sẽ giúp mô hình điều chỉnh sự tập trung lớn hơn ở các cặp từ (input, output) nếu chúng có vị trí tương đương hoặc gần tương đương. Với mô hình seq2seq thông thường thì việc mã hóa toàn bộ thông tin từ đầu vào (source) vào 1 vector cố định khiến mô hình khi thực hiện trên các câu dài (long sentence) không thực sự tốt, mặc dù sử dụng LSTM (BiLSTM, GRU [18]) để khắc phục điểm yếu của mạng RNN truyền thống với hiện tượng Gradient biến mất, nhưng như thế có vẻ vẫn chưa đủ, đặc biệt đối với những câu dài hơn những câu trong tập dữ liệu huấn luyện. Từ đó, trong bài báo, tác giả đề xuất 1 cơ chế cho phép mô hình có thể chú trọng vào những phần quan trọng và thay vì chỉ sử dụng véc tơ ngữ cảnh (context vector) được tạo ra từ lớp cuối cùng của bộ mã hóa, tác giả sử dụng tất cả các đầu ra của từng tế bào (cell) qua từng bước thời gian (timestep), kết hợp với trạng thái ẩn của từng tế bào để "tổng hợp" ra một véc tơ ngữ cảnh (attention vector) và dùng nó làm đầu vào cho từng tế bào trong bộ giải mã.



Hình 28. Cơ chế seq2seq phổ biến (không có attention) [33]



Hình 29. Cơ chế seq2seq có attention [33]

Nội dung cơ chế, với đầu vào có độ dài *n* và đầu ra có độ dài *m*, mỗi xác suất của kí tự đầu ra sẽ phụ thuộc vào kí tự trước đó , trạng thái ẩn của bộ giải mã và véc tơ ngữ cảnh theo công thức:

trong đó , là các tham số huấn luyện, trạng thái ẩn được tính thông qua mạng LSTM trong bộ giải mã với thông tin từ trạng thái ẩn trước đó , kí tự kết quả trước đó và véc tơ ngữ cảnh

Véc tơ ngữ cảnh được tính theo công thức:

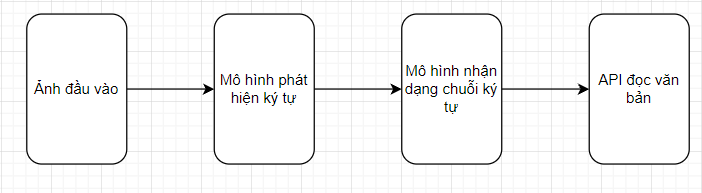
Với là trạng thái ẩn của bộ mã hóa, mỗi được gọi là điểm chú ý (attention score), được tính thông qua 1 hàm softmax là trọng số riêng được tính tại timestep *t* ứng với mỗi trạng thái ẩn của bộ mã hóa ​. Sử dụng hàm softmax như là một cách để chuẩn hóa đối với từng alignment score hay attention energy ​ thể hiện mức độ tương quan giữa và :

được gọi là mô hình căn chỉnh (alignment model) với mục đích đánh giá mức độ tương quan của của kí tự thứ *j* ở bộ mã hóa và kí tự thứ *i* ở bộ giải mã, có nhiều cách tính như Content-base Attention [19], Additive Attention [17], Multiplicative Attention [20], Dot Product [20],… Ở mô hình TBRA tác giả sử dụng Additive Attention:

trong đó , và là các ma trận trọng số.

## **2.2. Xây dựng mô hình**

Từ cơ sở lý thuyết ở trên, đồ án sẽ đề xuất mô hình hoàn chỉnh với nhiệm vụ chuyển đổi văn bản thành giọng nói, là sự kết hợp của những mô hình được huấn luyện riêng biệt.



Hình 30. Hệ thống chuyển đổi văn bản thành giọng nói

\* Phát hiện ký tự

Đầu vào của anh trong phạm vi đồ án được giả định đã căn chỉnh tương đối chuẩn xác với nền không phúc tạp. Mô hình CRAFT được huấn luyện trước của nhóm tác giả sẽ được áp dụng chop ha tiền xử lý. Hình ảnh đầu vào sau khi qua xử lý sẽ thu về kết quả là một danh sách các hộp giới hạn hình chữ nhật bao quanh mỗi từ, kết hợp với hàm để sắp xếp thứ tự các hộp đúng với thứ tự các chữ từ trái sang phải và theo từng dòng.

\* Nhận dạng chuỗi kí tự

Mô hình TBRA được sử dụng để dự đoán các từ từ ảnh đầu vào. Cụ thể như sau: đầu tiên những hình ảnh ở mức các từ được cắt từ các hộp giới hạn ở bước trước được đưa qua mô hình TRBA để thu được đầu ra có dạng một chuỗi ký tự s. Từ kết quả của ảnh đầu vào, một API sẽ được sử dụng để đọc nội dung dự đoán và phát cho người sử dụng.

## **2.3. Kết luận**

Chương 2 đã trình bày tổng quan về các phương pháp hiện có của từng bước trong bài toán nhận dạng và chuyển đổi văn bản thành giọng nói bao gồm các phương pháp tiền xử lý, phát hiện vùng kí tự và nhận dạng kí tự.

Ở chương tiếp theo, đồ án sẽ tiến hành xây dựng hệ thống chuyển đổi văn bản thành âm thanh áp dụng trong các thư viện đồng thời tiến huấn luyện các mô hình và thử nghiệm với hệ thống đã xây dựng.

# CHƯƠNG 3: ĐỀ XUẤT GIẢI PHÁP ỨNG DỤNG

## **1. Kịch bản hệ thống**

Đồ án đề xuất xây dựng hệ thống chuyển văn bản sách giấy và chuyển thành giọng nói để hỗ trợ người khiếm thị đọc sách tại các thư viện.

Kịch bản hệ thống gồm các bước như sau:

Bước 1: Huấn luyện các mô hình

Bước này được thực hiện trong lúc xây dựng hệ thống, các mô hình được huấn luyện từ những bộ dữ liệu văn bản tiếng Việt được thu thập trước đó hoặc tự sinh theo các tham số về font chữ, màu nền.

Bước 2: Trích xuất thông tin

Khi độc giả muốn đọc sách, nhân viên thư viện sẽ hỗ trợ thiết đặt sách vào vị trí phù hợp của camera, hình ảnh trang sách sẽ được chụp lại. Hình ảnh sẽ được xử lý để lấy các hộp giới hạn của từ và sắp xếp lại theo thứ tự lần lượt để tạo thành một đoạn văn có nghĩa. Sau đó hình ảnh và danh sách các hộp giới hạn được gửi sang server xử lý nhận diện

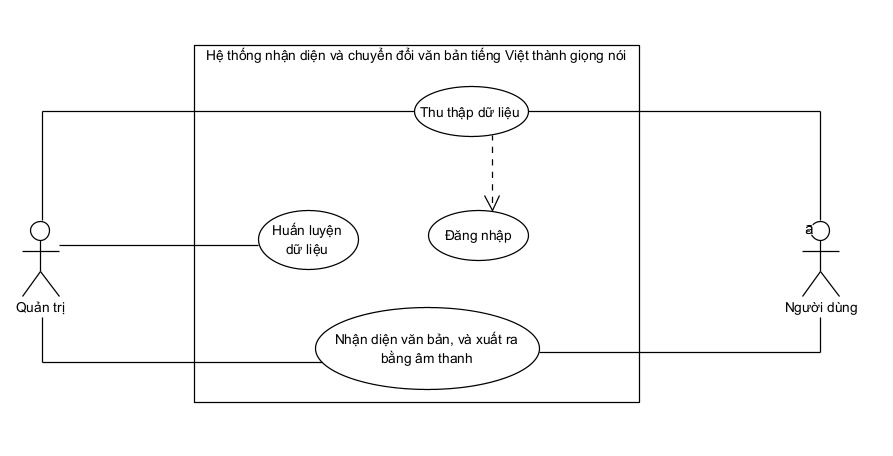
Bước 3: Nhận diện từ tiếng Việt

Với hình ảnh và toạ độ của các hộp giới hạn, server nhận diện sẽ dự đoán các từ Nội dung sau khi được nhận diện sẽ gửi lại và đọc thành tiếng cho độc giả.

## **2. Xây dựng hệ thống**

### **2.1. Phân tích hệ thống**

Từ kịch bản đề xuất ở trên, đồ án xây dựng hệ chuyển đổi văn bản thành giọng nói như hình dưới



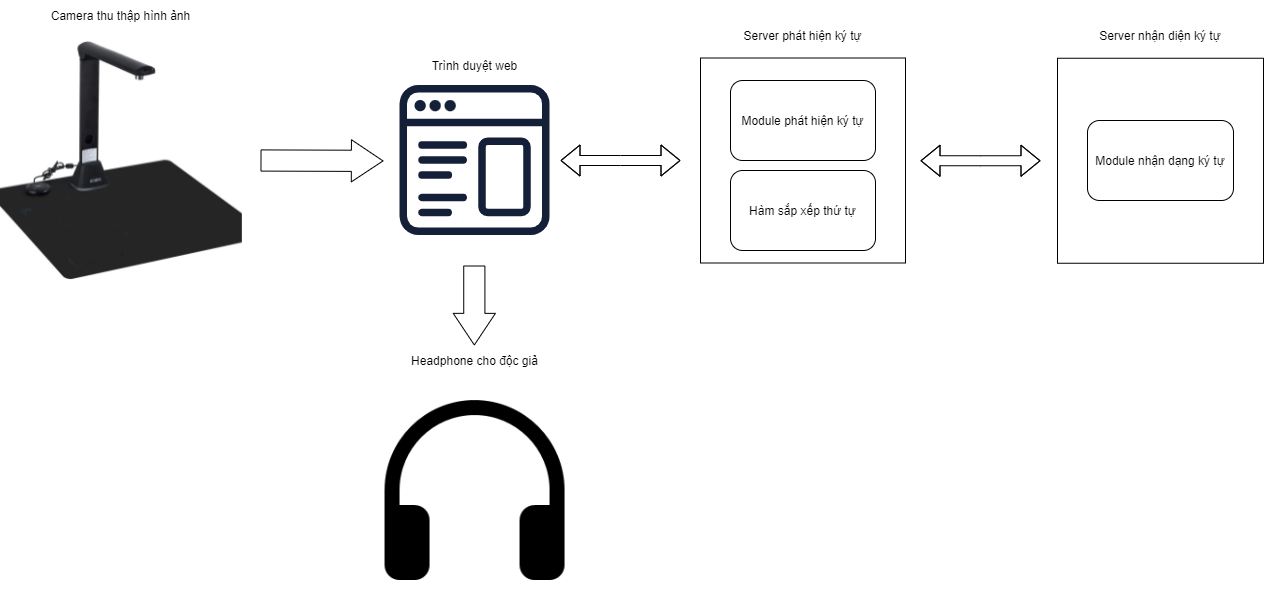
Hình 31. Biểu đồ ca sử dụng trong hệ thống

Hệ thống có 2 tác nhân chính:

* Người sử dụng: Bao gồm thủ thư và độc giả:
  + Thủ thư sẽ giúp đỡ độc giả chọn sách, đặt sách vào đúng vị trí phù hợp tại máy scan và hướng dẫn người đọc sử dụng (vị trí nút chụp hình, âm lượng…)
  + Độc giả sẽ bấm nút chụp và nghe nội dung qua tai nghe, khi kết thúc trang sách có thể tự mình lật trang và bấm nút chụp để tiếp tục nghe nội dung sách.
* Quản trị hệ thống: Thực hiện thu thập và huấn luyện cho model với nhiều ảnh đầu vào được người dùng đưa vào mỗi khi sử dụng.

### **2.2. Thiết kế hệ thống**

Từ kịch bản trên, đồ án đề xuất xây dựng hệ thống nhận dạng chuyển đổi văn bản thành giọng nói. Hệ thống dựa trên kiến trúc nhiều tầng, bao gồm 5 thành phần chính như hình 24.



Hình 32. Kiến trúc chung của hệ thống

* **Camera thu thập hình ảnh:** Thực hiện thu thập hình ảnh của trang sách.
* **Trình duyệt web:** Nhận ảnh được gửi từ người dùng, và gửi xử lý/trả kết quả dưới dạng tệp âm thanh phát trực tiếp.
* **Máy chủ phát hiện ký tự:**Là máy chủ với nhiệm vụ xử lý ảnh đầu vào, phát hiện các từ và sắp xếp lại thứ tự thành đoạn văn có nghĩa.
* **Máy chủ nhận dạng ký tự:** Là máy chủ cốt lõi của hệ thống, làm nhiệm vụ nhận ảnh từ máy chủ phát hiện ký tự, dự đoán các từ và trả về nội dung dự đoán của trang sách được đưa vào.

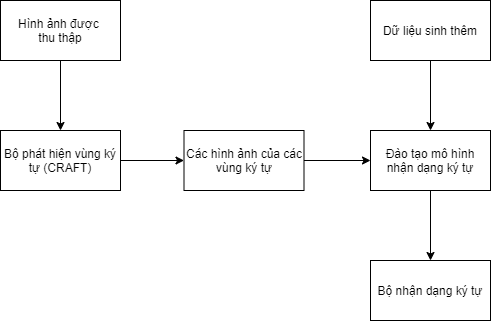
Các thành phần trên sẽ thực hiện 2 nhóm chức năng chính: Tiếp nhận/ trả thông tin và huấn luyện, nhận dạng. Mô đun Tiếp nhận/ trả thông tin chịu trách nhiệm các chức năng vào ra: Nhận ảnh từ camera, gửi xử lý và nhận xâu ký tự sau nhận diện để chuyển đổi thành giọng nói qua Google API. Còn module Huấn luyện và nhận dạng thực hiện các công việc chính liên quan đến thu thập, huấn luyện dữ liệu, nhận dạng văn bản… Trong phần tiếp theo, đồ án sẽ trình bày chi tiết mô đun Huấn luyện dữ liệu và Nhận dạng văn bản tiếng Việt.

## **3. Huấn luyện mô hình và nhận dạng trích xuất thông tin**

Phần này bao gồm hai quy trình được thực hiện theo vòng tuần hoàn. Mô hình nhận dạng trích xuất thông tin được huấn luyện trước với tập dữ liệu văn bản tiếng Việt được thu thập trước đó, kết hợp với bộ dự liệu được sinh theo thuật toán với các thông số cho trước về font chữ, nền ảnh.

### **3.1. Huấn luyện mô hình**

Việc huấn luyện các mô hình trong hệ thống bao gồm 2 pha như trên hình, được thực hiện bởi người quản trị. Hai pha này có thể thực hiện riêng biệt, không liên tục.

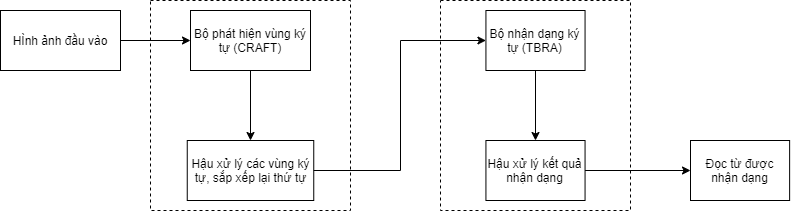


Hình 33. Quy trình huấn luyện mô hình

Quy trình thực hiện huấn luyện bộ nhận dạng kí tự, ảnh được đưa qua mô hình phát hiện vùng kí tự CRAFT, đã được huấn luyện trước (pretrained model) của nhóm tác giả với các tập dữ liệu gồm: tập dữ liệu sinh, IC13 và IC17, thu được ma trận điểm và ma trận ái lực. Với mục tiêu thu được kết quả là vùng kí tự ở mức từ ta thiết lập các tham số text\_threshold = 0.7 (ngưỡng xác định đó là văn bản), link\_threshold = 0.4 (ngưỡng xác định hai kí tự kết nối với nhau) để cho hộp giới hạn có thể bao quanh văn bản trong ảnh ở mức từ, các tham số còn lại được thiết lập như mặc định như trong mã nguồn của nhóm tác giả [28]. Sau đó thực hiện các bước hậu xử lý như phân ngưỡng, lọc nhiễu… như trong [28] để trả về kết quả là một danh sách các hộp giới hạn hình chữ nhật bao quanh mỗi từ. Để thu được các từ hợp lệ ta thực hiện các bước lọc như loại bỏ các hộp giới hạn với góc nghiêng lớn, cắt đôi những hộp giới hạn có chiều cao lớn hơn ngưỡng đã thiết lập, những trường hợp này xảy ra do có hai từ trên dưới dính vào nhau dẫn tới mô hình dự đoán đó là một từ. Các hộp giới hạn được tính toán lại để tạo thành các hộp giới hạn chữ nhật nằm ngang bằng cách xác định tọa độ của điểm trên cùng bên trái và điểm dưới cùng bên phải là giá trị nhỏ nhất và giá trị lớn nhất theo trục Ox Oy tương ứng. Các hộp giới hạn sau đó được cắt thành các hình ảnh của từ để phục vụ cho việc huấn luyện mô hình nhận dạng kí tự, mỗi hộp giới hạn trước khi cắt đều được mở rộng theo trục Oy 1 tỉ lệ để có thể bao phủ được các trường hợp hộp giới hạn không thể chứa được các dấu của từ.

### **3.2. Nhận dạng trích xuất thông tin**

Quá trình nhận dạng nhận dạng chuyển đổi văn bản thành giọng nói được thực hiện như mô hình đã xây dựng ở Chương 2. Tuy nhiên, với bài toán đã đặt ra ở đây không chỉ yêu cầu thuật toán có độ chính xác cao mà còn phải đảm bảo tính thời gian thực. Dựa trên các kết quả nghiên cứu thực nghiệm thông qua việc khảo sát, đánh giá các hướng tiếp cận tốt hiện có, đồ án đề xuất một giải pháp phù hợp, đạt hiệu quả cao (cho độ chính xác cao và xử lý thời gian thực) để giải quyết bài toán đã đặt ra.



Hình 34. Quy trình nhận dạng chuyển văn bản thành giọng nói của hệ thống

Các bước cơ bản để nhận diện văn bản tiếng Việt và chuyển thành giọng nói được mô tả như trên Hình 34. Cũng giống cơ bản với phần Huấn luyện mô hình nhưng được chia thành 2 pha tương ứng với từng phần phát hiện vùng kí tự, nhận dạng kí tự đã được đề cập. Cả 2 pha này sẽ được thực hiện trực tiếp, liên tục và thời gian thực, pha này nối tiếp pha kia.

Tại pha 1, hình ảnh trang sách sẽ được chụp và gửi lên hệ thống thông qua máy tính kết nối mạng. Hình ảnh sẽ được truyền vào bộ phận nhận diện ký tự của hệ thống để thu được toạ độ hộp giới hạn của các từ và sắp xếp lại theo thứ tự đúng theo từng dòng.

Tại pha 2, cũng tương tự như phần xác định hình ảnh các từ trong pha 2 ở phần Huấn luyện mô hình, hình ảnh gốc và thứ tự của các từ được xử lý để thu được những ảnh của từng từ, sau đó đưa vào bộ nhận dạng ký tự (TBRA) để dự đoán. Kết quả thu được sẽ được gửi trả lại giao diện và phát cho người nghe thông qua API của Google.

## **4. Thử nghiệm hệ thống**

Quy trình thử nghiệm hệ thống bao gồm hai quy trình: huấn luyện các mô hình con của mô hình nhận dạng ký tự và thiết lập thử nghiệm phần nhận dạng của hệ thống. Chi tiết của mỗi quy trình sẽ được đề cập ở phần sau.

### **4.1. Cài đặt huấn luyện các mô hình**

Bước này bao gồm việc huấn luyện các mô hình: mô hình phân đoạn U-Net, mô hình phân loại VGG16 và mô hình nhận dạng kí tự TBRA.

Việc cài đặthuấn luyện các mô hình được thực hiện trên môi trường máy ảo cung cấp bởi Google Colaboratory. Thông số cụ thể của môi trường được mô tả cụ thể trong bảng:

|  |  |
| --- | --- |
|  | Môi trường máy ảo của Google Colaboratory |
| Vi xử lý | Intel Xeon 2.30GHz 2 luồng |
| Dung lượng RAM | 13GB |
| Dung lượng bộ nhớ | 49 GB |
| GPU | Testla K80 12GB VRAM |
| Hệ điều hành | Ubuntu 18.04 |

Bảng 1. Môi trường cài đặt thử nghiệm việc huấn luyện các mô hình

Google Colaboratory là một môi trường Jupyter notebook miễn phí không yêu cầu cài đặt và hoạt động toàn bộ trên cloud. Jupyter Notebook là một ứng dụng chạy trên web cho phép người dùng tạo ra các văn bản, thường được gọi là notebook, bao gồm các đoạn code và đầu ra khi chạy chúng (có thể là chữ, hình ảnh, video, biểu đồ, v.v…). Jupyter Notebook hỗ trợ nhiều ngôn ngữ, bao gồm Python, R, Julia,… Google Colaboratory được xây dựng dựa trên Jupyter Notebook, tuy nhiên có điểm khác là cung cấp tài nguyên tính toán mạnh bao gồm: dung lượng RAM và bộ nhớ lớn, vi xử lý có hiệu năng tốt và GPU hỗ trợ tính toán. Tuy nhiên Google Colab hiện đang chỉ hỗ trợ ngôn ngữ Python. Các tài liệu chia sẻ trên Google Colaboratory đều là một notebook, nhưng được chạy trên cloud của Google Colaboratory và lưu trữ trong Google Drive tương ứng với tài khoản Google được sử dụng để tạo notebook. Môi trường của Colabotary có thể liên kết với Google Drive của người dùng để đọc và ghi các file từ đó. Môi trường có sẵn một số thư viện Python, bao gồm cà các thư viện dùng cho học máy phổ biến như TensorFlow hay scikỉt-learn.

***\* Mô hình nhận dạng kí tự TBRA***

\* Chuẩn bị dữ liệu

Như đã trình bày, đồ án huấn luyện mô hình TBRA sử dụng hai tập dữ liệu gán nhãn và tập dữ liệu sinh. Tập dữ liệu gán nhãn là những hình ảnh của các từ được cắt ra từ bước trước sau đó áp dụng mô hình có hỗ trợ sẵn tiếng Việt là Tesseract [31] để nhận dạng ký tự.

Tổng cộng lượng dữ liệu gán nhãn là 2 triệu hình ảnh bao gồm cả thực hiện tăng số lượng dữ liệu (Data Augmentation) bằng việc thêm các ảnh được làm mờ, thêm nhiễu, làm nghiêng,… Tập dữ liệu sinh là tập dữ liệu được tạo ra từ các công cụ sinh kí tự trên nền hình ảnh cho trước với các thông số về phông chữ, cỡ chữ, nền ảnh, độ nghiêng, độ nhiễu,… tùy chọn, sử dụng công cụ TextRecognitionDataGenerator [32]. Đồ án thực hiện tạo ra các từ văn bản ngẫu nhiên theo phông chữ, hình nền tương ứng với nhiều loại nền sách khác nhau. Dữ liệu cũng được làm mờ, thêm nhiễu, làm nghiêng để tăng sự đa dạng, tập dữ liệu sinh cũng gồm 2 triệu hình ảnh.

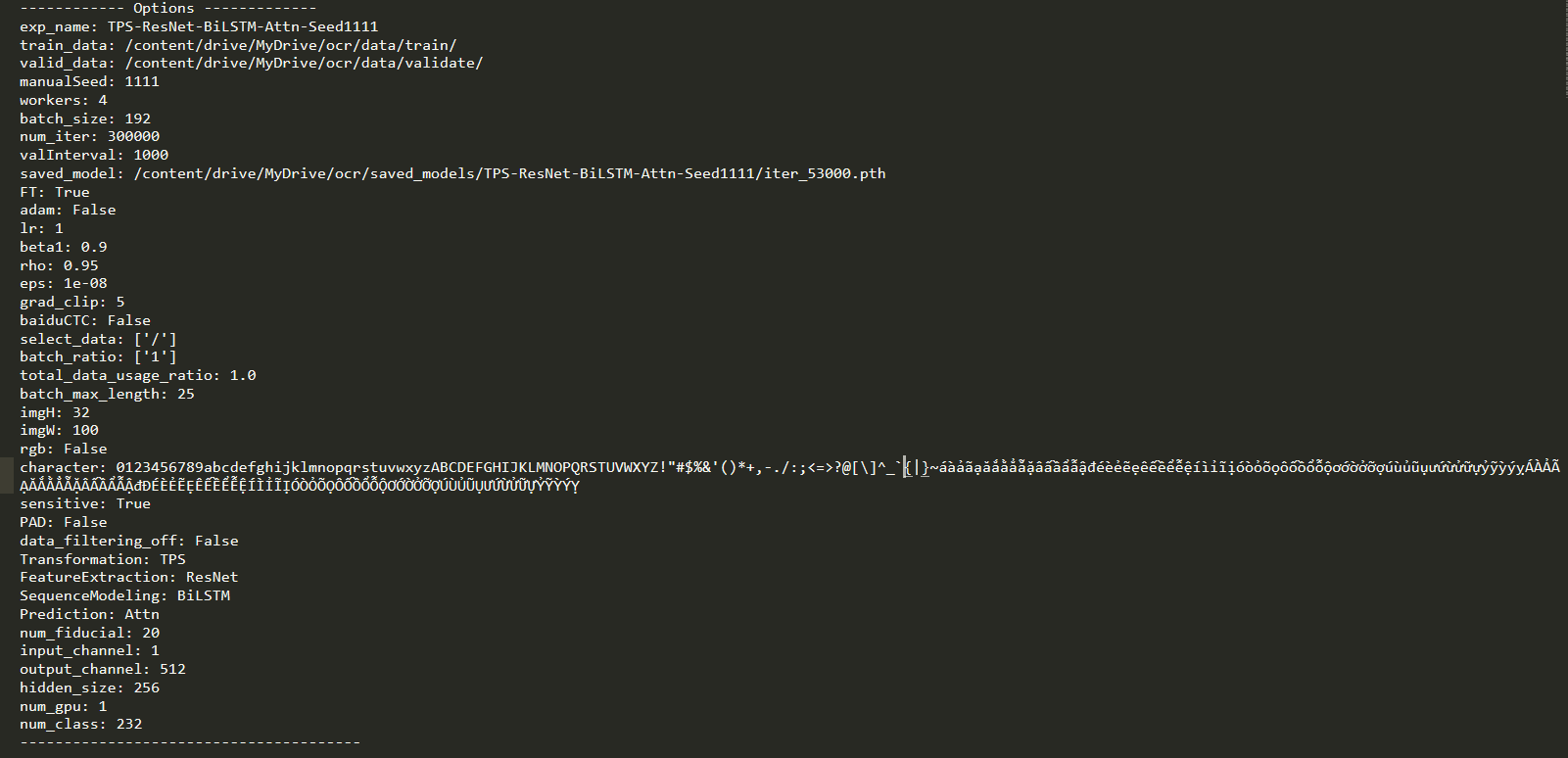
Với tập dữ liệu chia thành 2 tập con cho việc huấn luyện với tỉ lệ 80:20 trong đó 80% dữ liệu huấn luyện (training dataset) và 20% dữ liệu kiểm chứng (validation dataset). Toàn bộ dữ liệu được lưu trữ dưới dạng LMDB [30].

LMDB (Lightning Memory-Mapped Database), được phát triển và duy trì bởi Symas Corporation để thay thế [Berkeley DB](https://dbdb.io/db/berkeley-db) [27] trong dự án OpenLDAP, là một cơ sở dữ liệu nhúng có cấu trúc dạng khóa – giá trị và dựa trên cây B+. Các tính năng chính của LMDB là nó sử dụng một kho lưu trữ đơn cấp dựa trên các tệp bản đồ bộ nhớ, có nghĩa là hệ điều hành chịu trách nhiệm quản lý các trang (như bộ nhớ đệm thường xuyên sử dụng các trang). Nó sử dụng phương pháp lưu trữ copy-on-write với một luồng ghi duy nhất; những luồng đọc không chặn luồng ghi và ngược lại. Hệ thống chỉ duy trì nhiều nhất hai phiên bản dữ liệu vào bất kỳ lúc nào (tức là khi đã cam kết tất cả các phiên bản trước đó sẽ bị loại bỏ). Nó cũng duy trì một danh sách các trang trống để theo dõi và sử dụng lại các trang thay vì phân bổ bộ nhớ mỗi lần.

\* Thực hiện huấn luyện

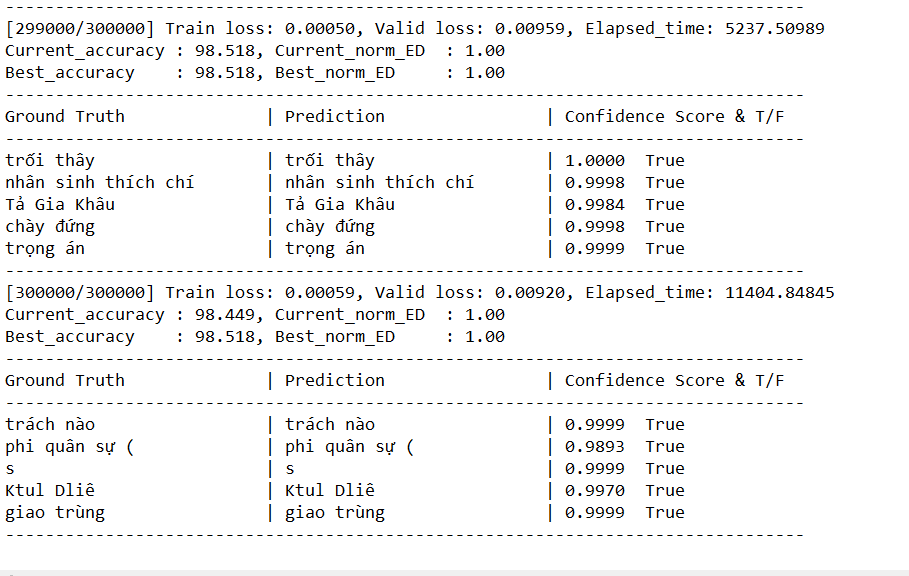
Việc huấn luyện được thực hiện với thư viện PyTorch, được phát triển bởi Facebook và là nền tảng tính toán khoa học được xây dựng để hỗ trợ chính cho các giải thuật học máy thường sử dụng GPU. Để được xem 1 thư viện Deep Learning lý tưởng, nó phải có tính dễ học và dễ sử dụng, đủ linh hoạt để sử dụng trong các ứng dụng khác nhau, hiệu quả để chúng ta có thể xử lý các bộ dữ liệu thực tế lớn và đủ chính xác để cung cấp kết quả chính xác ngay cả khi không chắc chắn về dữ liệu đầu vào. PyTorch được đánh giá là có thể đáp ứng rất tốt các yêu cầu nêu trên. Bên cạnh đó, PyTorch còn sỡ hữu khả năng tính toán với GPU, mang lại hiệu năng tính toán cao. Ngoài ra, phần lõi của PyTorch (libtorch) được viết bằng C++ giúp cài đặt các mô hình huấn luyện dễ dàng hơn khi phải thực hiện trên nhiều môi trường khác nhau.

Các đoạn mã huấn luyện được đồ án lấy từ mã nguồn của nhóm tác giả [29]. Theo đó, các tham số được thiết lập cho việc huấn luyện như trên hình 35.

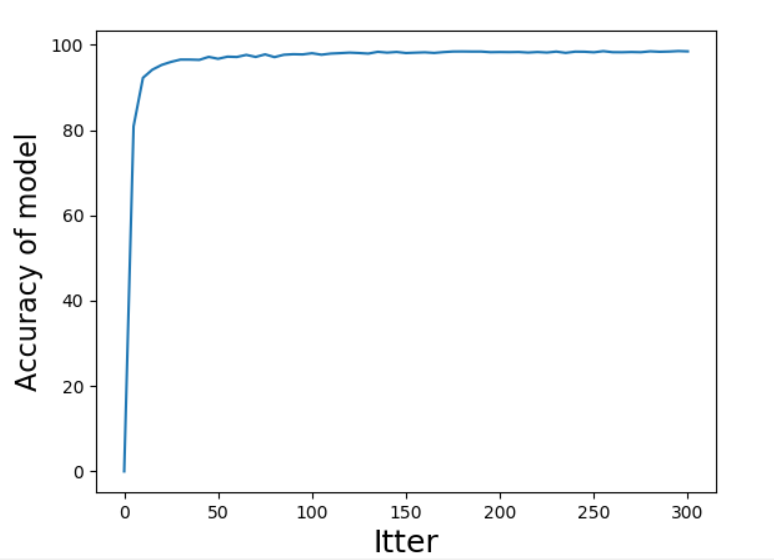


Hình 35. Tham số đào tạo TBRA

Trong đó select\_data, lựa chọn tập dữ liệu cho việc huấn luyện, thiết lập chọn toàn bộ tập dữ liệu huấn luyện. Batch\_ratio, tỉ lệ dữ liệu được chọn trong mỗi mẻ, được thiết lập là 1. Character, các nhãn kết quả của mỗi kí tự, được thiết lập là tất cả các kí tự trong tiếng Việt. Hình ảnh đầu vào là ảnh xám có kích thước 32x100 và batch\_size là 192. Mô hình dựa trên attention với việc so sánh các phân phối xác xuất đầu ra nên hàm mất mát là cross\_entropy, việc tối ưu hàm mất mát được thực hiện theo mặc định với hàm Adadelta [23]. Cuối cùng đồ án thực hiện huấn luyện bằng cách tinh chỉnh (fine turning) mô hình đã được đào tạo trước của nhóm tác giả với tập dữ liệu [MJSynth (MJ)](http://www.robots.ox.ac.uk/~vgg/data/text/) [1] và [SynthText (ST)](http://www.robots.ox.ac.uk/~vgg/data/scenetext/) [2].



Hình 36. Kết quả huấn luyện TBRA



Hình 37. Biểu đồ thể hiện độ hội tụ của accuracy

Kết quả huấn luyện được thể hiện như trên hình 34, với 300000 iteration và độ chính xác trên tập dữ liệu kiểm chứng đạt 98.518%.

### **4.2. Nhận dạng trích xuất thông tin**

Trong phần này, đồ án sẽ trình bày về các thiết lập thử nghiệm cho phần nhận dạng và chuyển đổi văn bản thành giọng nói, chi tiết việc cài đặt, kết quả thu được, đánh giá kết quả.

#### **4.2.1. Cài đặt các hệ thống**

Hệ thống được thiết lập trên máy tính cấu hình có cấu hình như sau:

|  |  |
| --- | --- |
| Hệ điều hành | Windows 10 |
| Vi xử lý | Intel core I5-7200U 2.50GHz 4 nhân 4 luồng |
| Dung lượng RAM | 8GB |
| Dung lượng bộ nhớ | 512 Gb |
| GPU | Intel HD Graphics 620 |

Bảng 2. Cấu hình hệ thống thử nghiệm

\* Cài đặt máy chủ ứng dụng

Máy chủ ứng dụng được cài đặt với ngôn ngữ Python phiên bản 3.6.

Các thư viện chính của hệ thống:

|  |  |  |
| --- | --- | --- |
| STT | Tên package | Phiên bản |
| 1 | numpy | 1.19.4 |
| 2 | tensorflow | 2.0.0 |
| 3 | Keras | 2.0.0 |
| 4 | torch | 1.5.0 |
| 6 | Flask | 1.1.2 |
| 7 | opencv-python | 3.4.2.16 |

Bảng 3.Các thư viện chính được sử dụng trong máy chủ hệ thống

Mỗi mô đun tiền xử lý, phát hiện vùng kí tự, nhận dạng kí tự trong máy chủ ứng dụng được xây dựng thành các RESTful API tương ứng với cổng 5000 cho máy chủ web và cổng 3010 cho máy chủ nhận diện, nhận yêu cầu từ máy chủ web và dữ liệu trao đổi có dạng JSON.

\* Cài đặt máy chủ web và trình duyệt phía người dùng

|  |  |
| --- | --- |
|  | Công nghệ, ứng dụng hỗ trợ |
| Frontend | - Thiết kế giao diện, xử lý dữ liệu, logic hiển thị: thư viện ReactJs, JavaScript.  - Tạo luồng stream: Socketio |
| Backend | - Ngôn ngữ: Python.  - Sử dụng framework Flask.  - Triển khai trên local host, cổng 5000 và 3010.  - Xây dựng các RESTful API nhận hình ảnh và các thông tin liên quan từ trình duyệt người gọi máy chủ ứng dụng để thực thi nhận dạng trích xuất thông tin. Bao gồm hai máy chủ phát hiện và nhận dạng. |

Bảng 4. Công nghệ sử dụng trong máy chủ web và trình duyệt phía người dùng

#### **4.2.2. Thiết lập thử nghiệm**

Mục đính chính của thử nghiệm bước nhận dạng trích xuất thông tin của hệ thống là để kiểm tra tính chính xác của hệ thống, bằng cách trả lời các câu hỏi, với một hình ảnh có chứa dữ liệu đầu vào liệu hệ thống:

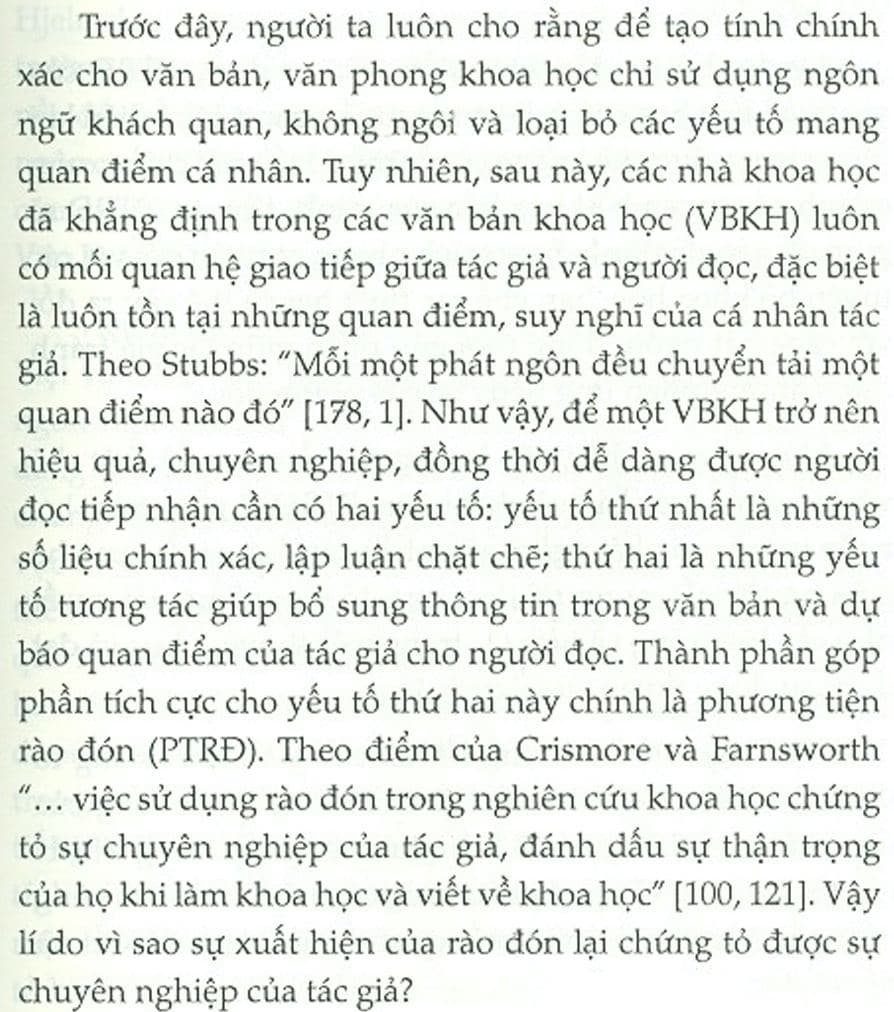
* Có thể phát hiện và sắp xếp đúng thứ tự các từ có trong ảnh hay không?
* Có thể nhận dạng đúng phần kí tự được cắt từ phần phát hiện hay không và việc hậu xử lý của phần này có tốt hay không?
* Có đáp ứng với thời gian cho phép trong việc xử lý một hình ảnh văn bản bất kỳ hay không?

Để thử nghiệm và đánh giá hiệu năng của hệ thống, đồ án thử nghiệm hệ thống với những hình ảnh trang sách không chỉ được chụp bởi thiết bị camera của hệ thống mà còn thử nghiệm với những hình ảnh khác được chụp ngẫu nhiên với các nền khác nhau và góc chụp và nền khác nhau.

Để đánh giá chất lượng của hệ thống, độ đo sau đây được sử dụng: độ chính xác (accuracy)

1. **Kịch bản 1: Ảnh đầu vào được đặt thẳng, background đều nhau**

Với kịch bản này, ứng dụng được thử nghiệm với ảnh đầu vào đạt mức độ tốt nhất có thể, khi trang sách được đặt thẳng, nền không bị loá hay có nhiễu, bóng.



Hình 38. Ảnh đầu vào với kịch bản 1

Với ảnh đầu vào như trên, hệ thống nhận diện đúng 213/145 từ, đạt accuracy 86,9%. Các từ được phát hiện đúng, đủ, không bị sai lệch thứ tự. Các trường hợp bị nhận diện sai do các lý do sau:

+ Các dầu câu được đặt liền với từ khiến cho hệ thống đưa ra dự đoán sai.

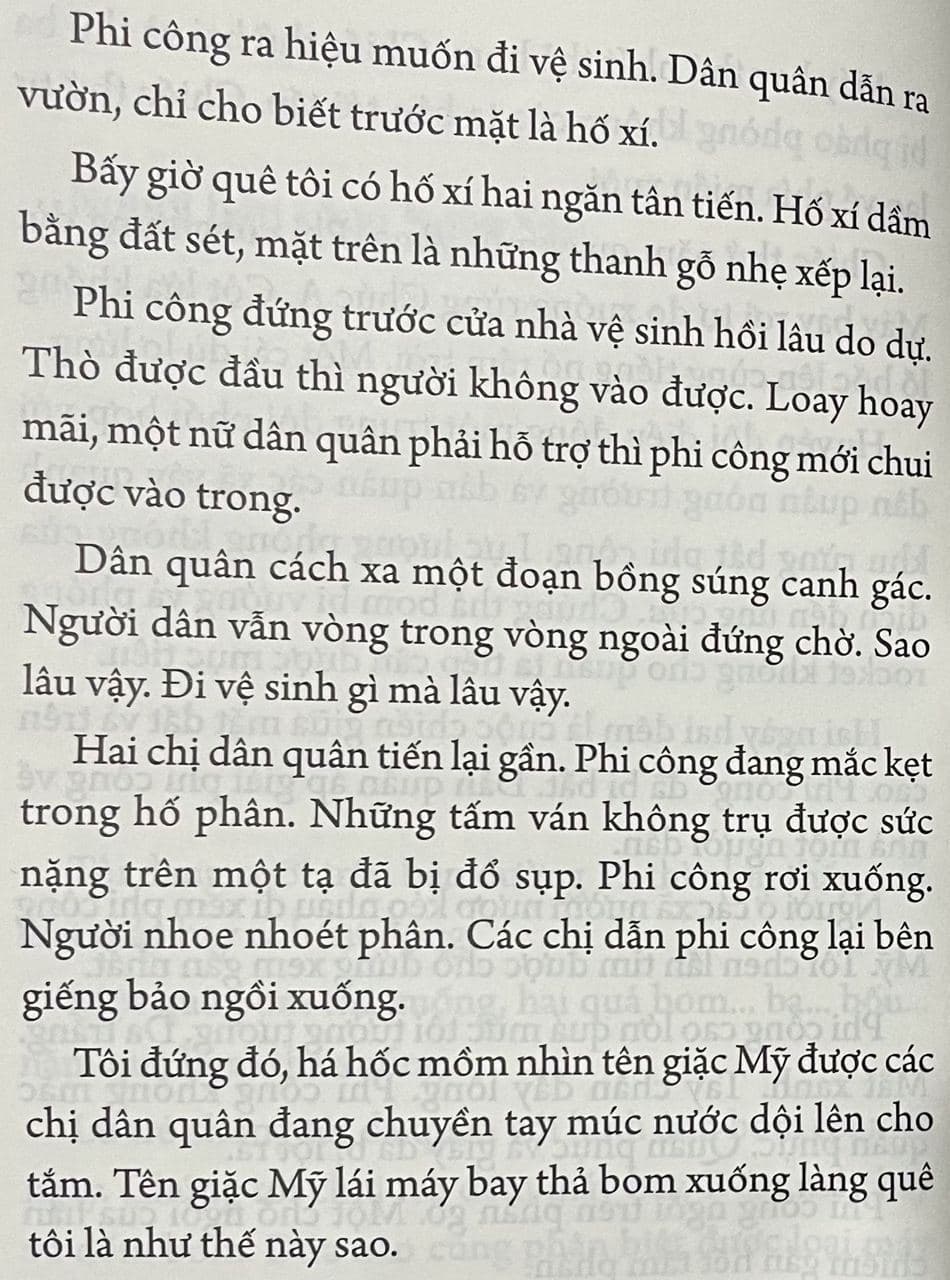
+ Phát hiện thiếu dấu của từ dẫn đến đưa ra dự đoán sai

|  |  |  |
| --- | --- | --- |
| Từ được phát hiện | Dự đoán của hệ thống | Confident score |
| D:\Study\Text-to-Speech\recog_server\static\out\227.png | Z OO. | 0.0780 |
| D:\Study\Text-to-Speech\recog_server\static\out\22.png | TìROI | 0.1995 |
| D:\Study\Text-to-Speech\recog_server\static\out\56.png | GVBKHề | 0.0749 |
| D:\Study\Text-to-Speech\recog_server\static\out\4.png | luỗn | 0.9102 |
| D:\Study\Text-to-Speech\recog_server\static\out\22.png | Ni3O> | 0.0312 |
| D:\Study\Text-to-Speech\recog_server\static\out\66.png | T Tlũ | 0.1164 |
| D:\Study\Text-to-Speech\recog_server\static\out\85.png | giảr | 0.3936 |
| D:\Study\Text-to-Speech\recog_server\static\out\99.png | đóế | 0.9340 |
| D:\Study\Text-to-Speech\recog_server\static\out\106.png | V KH | 0.4925 |
| D:\Study\Text-to-Speech\recog_server\static\out\145.png | g vếu | 0.2064 |

Bảng 5. Một số trường hợp nhận diện sai với kịch bản 1

1. **Kịch bản 2: Ảnh đầu vào được đặt nghiêng, background đều nhau**

Với kịch bản này, ứng dụng được thử nghiệm với ảnh đầu vào là trang sách bị nghiêng nhẹ, nền không bị loá hay bóng, tuy nhiên có nội dung của trang sau in hằn lên.

****

Hình 39. Ảnh đầu vào với kịch bản 2

Với ảnh đầu vào như trên, hệ thống nhận diện đúng 165/210 từ, đạt accuracy 75,5%. Các lỗi sai từ hệ thống đối với kịch bản này bao gồm:

+ Phát hiện sai từ, xác định những phần giấy bị in hằn là một từ.

+ Thứ tự của từ trong đoạn văn chưa được sắp xếp đúng.

+ Các dầu câu được đặt liền với từ khiến cho hệ thống đưa ra dự đoán sai.

+ Dự đoán còn nhầm lẫn giữa các dấu câu “hỏi”, “ngã” và dấu mũ

|  |  |  |
| --- | --- | --- |
| Từ được phát hiện | Dự đoán của hệ thống | Confident score |
| D:\Study\Text-to-Speech\recog_server\static\out\188.png | , | 0.0815 |
| D:\Study\Text-to-Speech\recog_server\static\out\205.png | quẫn | 0.5017 |
| D:\Study\Text-to-Speech\recog_server\static\out\8.png | vườm, | 0.5744 |
| D:\Study\Text-to-Speech\recog_server\static\out\2.png | Tâ | 0.7960 |
| D:\Study\Text-to-Speech\recog_server\static\out\12.png | dãn | 0.9991 |
| D:\Study\Text-to-Speech\recog_server\static\out\40.png | dâm | 0.9976 |
| D:\Study\Text-to-Speech\recog_server\static\out\43.png | nhỡng | 0.9960 |
| D:\Study\Text-to-Speech\recog_server\static\out\50.png | eông | 0.6805 |
| D:\Study\Text-to-Speech\recog_server\static\out\57.png | hôi | 1.0000 |
| D:\Study\Text-to-Speech\recog_server\static\out\128.png | ỏ | 0.3307 |
| D:\Study\Text-to-Speech\recog_server\static\out\127.png | t ó | 0.4513 |
| D:\Study\Text-to-Speech\recog_server\static\out\129.png | c è | 0.2771 |

Bảng 6. Một số trường hợp nhận diện sai với kịch bản 2

1. **Kịch bản 3: Ảnh đầu vào bị loá**

Với kịch bản này, ứng dụng được thử nghiệm với ảnh đầu vào là trang sách có nền bóng, dẫn đến bị loá dưới ánh sáng đèn



Hình 40. Ảnh đầu vào với kịch bản 3

Với ảnh đầu vào như trên, hệ thống gần như không thể phát hiện đúng các từ dẫn đến các dự đoán đều sai, các kết quả phát hiện như bảng sau:

|  |  |  |
| --- | --- | --- |
| Từ được phát hiện | Dự đoán của hệ thống | Confident score |
| D:\Study\Text-to-Speech\recog_server\static\out\0.png | r-eo-ne | 0.0870 |
| D:\Study\Text-to-Speech\recog_server\static\out\4.png | è n.ee | 0.1393 |
| D:\Study\Text-to-Speech\recog_server\static\out\15.png | easee | 0.5744 |

Bảng 7. Một số trường hợp nhận diện sai với kịch bản 2

#### **4.2.3. Đánh giá kết quả**

Kết quả thử nghiệm với 20 hình ảnh trang sách ngẫu nhiên như sau:

|  |  |
| --- | --- |
| Các trường hợp | Kết quả (số từ) |
| Nội dung đúng, không tính trường hợp thông tin mờ | 1738 |
| Phát hiện sai vùng kí tự dẫn đến kết quả sai | 80 |
| Hậu xử lý vùng kí tự dẫn đến kết quả sai (bị cắt vào chữ, sắp xếp sai thứ tự các từ) | 476 |
| Phát hiện kí tự đúng nhưng hậu xử lý sai dẫn đến kết quả sai | 847 |
| Không phát hiện được vùng kí tư | 53 |

Bảng 8. Kết quả thử nghiệm hệ thống

Từ kết quả thử nghiệm trên có thể kết luận như sau: Hệ thống trích xuất đúng thông tin các trường khá chính xác trong điều kiện ảnh đầu vào được sắp thẳng với nền không quá phức tạp; Thời gian trung bình để trích xuất một ảnh và đọc nội dung nằm trong khoảng 15 – 20 giây. Với những ảnh nhận diện sai, nguyên nhân do nền ảnh phức tạp, hoặc có chữ in hằn từ trang sau lên trang trước khiến vùng không có ký tự cũng được nhận dạng thành ký tự. Về phần phát hiện vùng kí tự và nhận dạng, những trường hợp phát hiện vùng kí tự sai gồm vùng của từ này đè lên từ khác (chữ của trang sách sau in lên trang sách được chụp) hoặc xác định nền phức tạp như bị loá đèn, hoặc trang sách bị cong quá nhiều. Và giải pháp cần là thực hiện tinh chỉnh mô hình đào tạo trước CRAFT với tập dữ liệu chỉ có ảnh với nhiều góc độ hơn. Về mô hình dự đoán, những từ được phát hiện đúng nhưng cho kết quả nhận diện sai nguyên nhân do model chưa được đào tạo với tập dữ liệu đủ lớn, dẫn đến những từ nghiêng hoặc cong theo trang sách hoặc loá đèn chưa mang đến kết quả tốt. Bên cạnh đó, còn tồn tại tình trạng nhầm lẫn giữa các dấu câu của tiếng Việt, dẫn đến phát mang tính chất giọng địa phương, ngọng và sai chính tả. Cách khắc phục là gia tăng tập dữ liệu huấn luyện với đa dạng phông chữ, đặc biệt là phân biệt giữa các từ có dấu tương tự nhau; Cùng với đó là cải thiện thuật toán giúp xoay ảnh, sắp xếp từ theo phương thẳng để tăng hiệu năng nhận diện cũng như phân biệt tốt hơn đối với ảnh có nền phức tạp như có chữ bị in hằn hay loá do ánh sáng.

## **5. Kết luận**

Chương này trình bày quy trình đồ án áp dụng để thử nghiệm và đánh giá hệ thống. Do dữ liệu ít nên chỉ thử nghiệm đánh giá trên 300 bộ dữ liệu mẫu. Từ kết quả thu được, có thể rút ra kết luận hệ thống hoạt động ổn định với các trường hợp ảnh được chụp từ thiết bị trong hệ thống rõ ràng, điều kiện ánh phù hợp. Còn với các trường hợp ảnh mờ, bị nhòe màu do để thời gian lâu, các trường in lộn xộn vị trí thì hệ thống có thể nhận dạng nhầm hoặc sai.

# CHƯƠNG 4: KẾT LUẬN

## **1. Đóng góp của đồ án**

Với sự phát triển của các phương pháp học sâu, độ chính xác của các bài toán phân đoạn, phát hiện, nhận dạng ngày càng cao. Một mặt nhờ các giải thuật tiên tiến, một mặt nhờ

Đồ án này đã tập trung nghiên cứu công nghệ nhận dạng kí tự quang học và xây dựng hệ thống chuyển đổi nội dung văn bản thành giọng nói. Kết quả đạt được của đồ án bao gồm:

- Làm chủ quy trình, các bước để xây dựng một hệ thống áp dụng công nghệ nhận dạng kí tự quang học.

- Làm chủ được giải pháp công nghệ và các kỹ thuật hiện đại để giải quyết bài toán nhận dạng kí tự quang học, cụ thể bao gồm:

+ Các phương pháp phân đoạn, phát hiện hình ảnh, phương pháp phát hiện các vùng kí tự trong hình ảnh và nhận dạng kí tự áp dụng mạng neuron học sâu.

+ Các phương pháp xử lý hình ảnh số, các giải thuật để phục vụ trong việc xử lý các chuỗi kí tự.

- Nắm được việc xây dựng cơ bản hệ thống bao gồm cả Frontend và Backend với các ngôn ngữ Javascript và Python.

- Hệ thống chuyển đổi hình ảnh văn bản thành giọng nói đạt độ chính xác cao.

## **2. Hướng phát triển tiếp theo**

Trong quá trình triển khai, khó khăn lớn nhất đồ án gặp phải đó là quá trình thu thập dữ liệu, vì số lượng ảnh nhân dân lớn, dữ liệu thu thập lại thường bị ảnh hưởng do điều kiện môi trường (ánh sáng, nhiễu …). Nên trong tương lai nhóm tác giả sẽ nghiên cứu, phát triển công cụ dựa trên mạng xã hộ để thu thập dữ liệu ảnh tự động và bán tự động. Tiếp nữa đồ án muốn phát phiển hệ thống có khả năng trích xuất thông tin của nhiều loại giấy tờ hơn như bằng lái xe, hộ chiếu hay chứng minh thư nhân dân.

TÀI LIỆU THAM KHẢO

[1]: O. Ronneberger, P. Fischer, and T. Brox. “U-net: Convolutional networks for biomedical image segmentation”. In MICCAI, 2015.

[2]: Y. Bengio, P. Y. Simard, and P. Frasconi (1994). “Learning longterm dependencies with gradient descent is difficult”.

[3]: A. Graves, A. Mohamed, and G. E. Hinton. “Speech recognition with deep recurrent neural networks”. In ICASSP, 2013.

[4]: K. Simonyan and A. Zisserman. “Very deep convolutional networks for large-scale image recognition”. In ICLR, 2015.

[5]: Baoguang Shi, Xiang Bai, and Cong Yao. “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition”. In TPAMI, volume 39, pages 2298–2304. IEEE, 2017

[6]: M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. “Reading text in the wild with convolutional neural networks”. IJCV (Accepted), 2015

[7]: Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. “Spatial transformer networks”. In NIPS, pages 2017–2025, 2015

[8]: Chen-Yu Lee and Simon Osindero. “Recursive recurrent nets with attention modeling for ocr in the wild”. In CVPR, pages 2231–2239, 2016

[9]: A. Graves, S. Fernandez, F. J. Gomez, and J. Schmidhu-ber. “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks”. In ICML, 2006.

[10]: Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. “Robust scene text recognition with automatic rectification”. In CVPR, pages 4168–4176, 2016.

[11]: S. Hochreiter and J. Schmidhuber. “Long short-term memory”. Neural Computation, 9(8):1735–1780, 1997.

[12]: Baek, Jeonghun et al. “What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis.” 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019): 4714-4722.

[13]: Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In CVPR, pages 770–778, 2016.

[14]: Wei Liu, Chaofeng Chen, Kwan-Yee K Wong, Zhizhong Su, and Junyu Han. “Star-net: A spatial attention residue network for scene text recognition”. In BMVC, volume 2, 2016.

[15]: Jianfeng Wang and Xiaolin Hu. “Gated recurrent convolution neural network for ocr”. In NIPS, pages 334–343, 2017.

[16]: Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. “Rosetta: Large scale system for text detection and recognition in images”. In KDD, pages 71–79, 2018.

[17]: Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”. In ICLR, 2015.

[18]: K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. “On the properties of neural machine translation: Encoder-decoder approaches”. CoRR, abs/1409.1259, 2014.

[19]: A. Graves, G. Wayne, I. Danihelka (2014). "Neural Turing Machines". In arXiv:1410.5401.

[20]: Minh-Thang Luong, Hieu Pham, Christopher D. Manning (2015). “Effective Approaches to Attention-based Neural Machine Translation”. In arXiv:1508.04025.

[21]: https://cvit.iiit.ac.in/research/projects/cvit-projects/the-iiit-5k-word-dataset

[22]: https://rrc.cvc.uab.es/

[23]: https://d2l.aivivn.com/chapter\_convolutional-modern/resnet\_vn.html

[24]: https://towardsdatascience.com/everything-you-ever-wanted-to-know-about-computer-vision-heres-a-look-why-it-s-so-awesome-e8a58dfb641e

[25]: https://innotech-vn.com/giai-phap-ocr-do-chinh-xac-cao-cho-ngan-hang-tai-viet-nam/

[26]: http://csnd.vn/Home/Nghien-cuu-Trao-doi/2503/Tim-hieu-ve-cong-nghe-nhan-dang-ki-tu-quang-hoc-OCR-Optical-Character-Recognition-ung-dung-trong-cac-phan-mem-phat-hien-va-xu-ly-vi-pham-luat-giao-thong-thong-qua-he-thong-camera-giam-sat.

[27]: https://www.oracle.com/database/technologies/related/berkeleydb.html

[28]: https://github.com/clovaai/CRAFT-pytorch

[29]: https://github.com/clovaai/deep-text-recognition-benchmark

[30]: https://lmdb.readthedocs.io/en/release/

[31]: https://github.com/tesseract-ocr/tesseract

[32]: https://github.com/Belval/TextRecognitionDataGenerator

[33]: https://viblo.asia/p/machine-learning-attention-attention-attention-eW65GPJYKDO

[34]: https://towardsdatascience.com/understanding-semantic-segmentation-with-unet-6be4f42d4b47

[35]: https://arxiv.org/pdf/2106.15561.pdf

[36]: https://www.pyimagesearch.com/2021/04/17/image-classification-basics/

[37]: https://datalya.com/blog/machine-learning/image-classification-with-localization

[38]: https://www.fritz.ai/object-detection/

[39]: https://phamdinhkhanh.github.io/2020/06/10/ImageSegmention.html#12-bài-toán-image-segmentation

[40]: https://arxiv.org/ftp/arxiv/papers/1003/1003.3654.pdf

[41]: https://arxiv.org/pdf/1904.01941.pdf

[42]: https://arxiv.org/pdf/1507.05717.pdf

[43]: https://towardsdatascience.com/the-vanishing-gradient-problem-69bf08b15484

[44]: https://nttuan8.com/bai-14-long-short-term-memory-lstm/