

	VIETTEL AI RACE	TD005
	ỨNG DỤNG CÔNG NGHỆ NHẬN DẠNG KÝ TỰ QUANG HỌC	Lần ban hành: 1

Chuyển đổi số các cơ sở giáo dục đào tạo đang là một nhiệm vụ cấp bách và Học viện Ngân hàng cũng không nằm ngoài guồng quay của cuộc cách mạng công nghệ này. Để quá trình chuyển đổi số được thuận lợi thì quá trình số hóa dữ liệu cần được ưu tiên đẩy mạnh. Trong bài báo này, tác giả nghiên cứu những giải pháp phù hợp cho việc số hóa dữ liệu văn bản đã và đang lưu hành trong Học viện Ngân hàng. Cụ thể, tác giả giới thiệu các công nghệ cốt lõi trong số hóa tài liệu như Nhận dạng ký tự quang học, Xử lý văn bản thông minh, khảo sát các giải pháp tiêu biểu trên thị trường ở Việt Nam để lựa chọn giải pháp khả thi và tiến hành cài đặt thử nghiệm giải pháp FPT. AI Reader với bộ dữ liệu tự thu thập từ một số phòng ban tại Học viện. Kết quả thực nghiệm cho thấy tỷ lệ sai số ở mức từ đạt 27% và chỉ 16% từ bị sai ở các tiêu đề, đơn vị ban hành, loại văn bản. Giải pháp hoàn toàn có thể được nghiên

1. Nội dung chính

1.1 Đặt vấn đề

Thực hiện chuyển đổi số (CDS) trong lĩnh vực giáo dục đào tạo là một trong những hoạt động nhận được nhiều sự quan tâm và ưu tiên của nhiều nước trên thế giới. Hoạt động CDS trong lĩnh vực giáo dục không chỉ tập trung vào hoạt động dạy và học trong thay đổi phương pháp dạy và học mà còn diễn ra ở rất nhiều nghiệp vụ khác như phát triển hệ thống hỗ trợ hoạt động quản lý điều hành, quản lý hoạt động khoa học công nghệ. Trên cơ sở Quyết định số 131/ QĐ-TTg năm 2022 của Thủ tướng Chính phủ phê duyệt Đề án “Tăng cường ứng dụng công nghệ thông tin và chuyển đổi số trong giáo dục và đào tạo giai đoạn 2022-2025, định hướng đến năm 2030” (Thủ tướng Chính phủ, 2022), Học viện Ngân hàng cũng đã có những hành động thiết thực để thích ứng với bối cảnh phát triển chung của giáo dục đại học.

Liên quan đến chuyển đổi số, khá nhiều thuật ngữ với hậu tố “số” có thể gây nhầm lẫn, như “công nghệ số”, “kinh tế số”, “kỹ thuật số”. Đặc biệt thuật ngữ tiếng Anh cũng có những thuật ngữ gần giống nhau như Digitization, Digitalization, vậy vai trò của chúng đối với CDS như thế nào? Theo Phạm Huy Giao (2020) quá trình CDS bao gồm ba giai đoạn như Hình 1.

Theo đó, một tổ chức muốn thực hiện CDS, trước hết cần trải qua giai đoạn đầu tiên: Số hóa (Digitization). Đây là quá trình chuyển đổi các thực thể trong quy trình hoạt động từ dạng vật lý sang dạng số, có thể lưu trữ và xử lý trên máy tính điện tử. Chẳng hạn hồ sơ của một nhân viên từ bản sơ yếu lý lịch trên giấy được lưu trữ thành các trường thông tin

	VIETTEL AI RACE	TD005
	ỨNG DỤNG CÔNG NGHỆ NHẬN DẠNG KÝ TỰ QUANG HỌC	Lần ban hành: 1

trong cơ sở dữ liệu như Họ tên, Ngày sinh, Quê quán. Các số liệu kinh doanh từ việc lưu cả tập hóa đơn, chứng từ và kiểm kê lại khi cần thì được tổ chức thành các bảng số liệu trong Microsoft Excel. Hay chỉ đơn giản là lưu trữ các tệp tin (file) trong máy tính chứa các dữ liệu hoạt động của tổ chức cũng có thể phần nào được coi là số hóa.

Tại Học viện Ngân hàng, một số phòng ban đã có phần mềm quản lý và vận hành khá ổn định. Chẳng hạn Phòng Đào tạo có phần mềm Quản lý đào tạo, phòng Quản lý người học thực hiện quản lý trên Cổng thông tin sinh viên, Phòng Quản trị có phần mềm MISA, Tạp chí Khoa học và Đào tạo Ngân hàng cũng sử dụng phần mềm quản lý bài viết trực tuyến. Điều này giúp một lượng lớn dữ liệu tác nghiệp của Học viện được chuyển hóa vào các cơ sở dữ liệu tương ứng. Tuy nhiên, theo quan sát thực tế vẫn còn không ít dữ liệu vẫn đang lưu hành bản cứng, như các quyết định, tờ trình cũ khi chưa áp dụng phần mềm quản lý, hoặc các đơn từ, đặc biệt là đơn từ của sinh viên thì chưa có hình thức lưu trữ điện tử phù hợp. Khi cần phải tra cứu chỉ có thể đến tận nơi và tìm theo các tủ hồ sơ vật lý hoặc tra cứu trong máy tính cục bộ của chuyên viên phụ trách. Một số tài liệu được đăng tải trên website của đơn vị, nhưng chỉ có thể tra cứu theo tiêu đề văn bản (nếu được đặt tiêu đề chính xác) chứ các nội dung thường được đóng gói trong mã nhúng file PDF chỉ cho phép đọc chứ không thể truy cập chi tiết. Đây là một thách thức trong quá trình số hóa nói riêng và chuyển đổi số nói chung tại Học viện, thôi thúc các nghiên cứu tìm kiếm giải pháp thích hợp nhằm biến đổi các văn bản thô còn tồn đọng thành những tài liệu định dạng kỹ thuật số.

Từ thực trạng trên, tác giả đặt ra ba câu hỏi nghiên cứu chính:

- Những công nghệ cần thiết cho việc số hóa dữ liệu văn bản là gì?
- Có những giải pháp nào trên nền tảng những công nghệ đó thích hợp với số hóa dữ liệu văn bản tại Việt Nam?
- Những văn bản đã và đang lưu hành tại Học viện khi áp dụng thử nghiệm các giải pháp này cho kết quả như thế nào?

Để tìm kiếm lời giải cho các câu hỏi trên, tác giả đề xuất nghiên cứu và đánh giá các giải pháp và công nghệ hiện đại liên quan đến việc số hóa tài liệu, văn bản. Cụ thể tác giả đặt ra một số mục tiêu nghiên cứu sau:

Nghiên cứu tổng quan các công nghệ số hóa dữ liệu văn bản như nhận dạng ký tự quang học (OCR) hay xử lý văn bản thông minh (IDP). Đây đều là những công nghệ đóng vai trò rất quan trọng trong công cuộc chuyển đổi số.

	VIETTEL AI RACE	TD005
	ỨNG DỤNG CÔNG NGHỆ NHẬN DẠNG KÝ TỰ QUANG HỌC	Lần ban hành: 1

Khảo sát các giải pháp sẵn có trên thị trường ở Việt Nam, lựa chọn giải pháp khả thi cũng như cài đặt thử nghiệm với dữ liệu thực tế tại Học viện.

Thu thập dữ liệu thực tế từ các phòng ban chức năng và khoa chuyên ngành của Học viện. Dữ liệu được thu thập đảm bảo sự đa dạng về thể loại (các loại văn bản khác nhau lưu hành ở Học viện như nghị quyết, quyết định, thông báo), nguồn gốc (đơn vị phụ trách biên soạn hoặc ban hành văn bản), định dạng tài liệu (hình ảnh chụp từ điện thoại, ảnh quét từ máy scan, file pdf được chuyển từ văn bản MS Word).

Đánh giá mức độ hiệu quả về khả năng bóc tách tài liệu và nhận dạng ký tự của giải pháp đề xuất trên bộ dữ liệu đã thu thập.

Để đạt mục tiêu nghiên cứu, tác giả thực hiện tổng quan nhằm làm rõ cơ sở lý thuyết về công nghệ nhận dạng ký tự quang học; quan sát và trải nghiệm thực tế nhằm hệ thống các giải pháp nhận dạng ký tự quang học tại Việt Nam; và thực hiện thử nghiệm với tài liệu tại Học viện Ngân hàng.

1.2 Cơ sở lý thuyết về công nghệ nhận dạng ký tự quang học

Công nghệ Nhận dạng ký tự quang học-Optical Character Recognition (OCR) là một loại công nghệ cho phép máy tính điện tử tự động nhận biết các ký tự (chữ cái, số, dấu câu, ký tự đặc biệt) trên những hình ảnh được cung cấp (Ravina Mithe, 2013). Không giống như bộ não con người, thứ có thể dễ dàng đọc được các ký tự, câu chữ từ hình ảnh, máy tính không đủ thông minh và khả năng trừu tượng để nhận biết được loại thông tin này. Máy tính chỉ hiểu hình ảnh là các điểm ảnh (pixel) đại diện bởi các con số chỉ mã màu sắc ở pixel đó. Bởi vậy, nghiên cứu về công nghệ OCR vẫn đang là một chủ đề rất được quan tâm trong cộng đồng nghiên cứu Trí tuệ nhân tạo.

Cách thức hoạt động chung của OCR được mô tả ở Hình 2. Theo đó máy quét sẽ quét biểu mẫu chứa hình ảnh ký tự, sau đó công cụ nhận dạng tiến hành đọc hiểu các hình ảnh và chuyển chúng thành dữ liệu ASCII (các ký tự máy có thể đọc được). Có nhiều yếu tố ảnh hưởng đến chất lượng văn bản đầu ra của hệ thống OCR như chất lượng hình ảnh đầu vào (độ phân giải cao/thấp, góc chụp nghiêng/thẳng, độ sáng, độ bóng...), mật độ văn bản trên hình ảnh đầu vào (ví dụ giấy tờ cá nhân thì mật độ văn bản ít hơn so với các quy định pháp luật), phong chữ của tài liệu gốc (chữ viết tay, chữ in hoa, loại phong chữ) hay ngôn ngữ của tài liệu gốc (tiếng Việt, tiếng Anh hay nhiều ngôn ngữ cùng trong một văn bản).

	VIETTEL AI RACE	TD005
	ỨNG DỤNG CÔNG NGHỆ NHẬN DẠNG KÝ TỰ QUANG HỌC	Lần ban hành: 1

Quá trình xử lý của hệ thống OCR thường được chia thành 3 giai đoạn: phát hiện văn bản (text detection), nhận dạng văn bản (text recognition) và trích xuất thông tin (information extraction). Các giai đoạn này lại áp dụng nhiều kỹ thuật, thuật toán phức tạp của thị giác máy tính (Computer Vision) và xử lý ngôn ngữ tự nhiên (Natural Language Processing). Chẳng hạn với riêng bài toán trích xuất thông tin từ hóa đơn bán hàng tiếng Việt trong cuộc thi MC-OCR Challenge tổ chức năm 2021, mô hình nhận dạng ký tự tốt nhất được xây dựng dựa trên mạng nơ ron Faster R-CNN kết hợp với mạng nơ ron TransformerOCR (Vu Xuan-Son, 2021).

Cùng với OCR, gần đây, thuật ngữ Nhận dạng ký tự thông minh (Intelligent Character Recognition- ICR) (Raymond Ptucha, 2019) được sử dụng để mô tả quá trình đọc hiểu dữ liệu hình ảnh, cụ thể là văn bản chữ và số. ICR là một mô-đun của OCR, có khả năng biến hình ảnh viết tay hoặc các ký tự in thành dữ liệu ASCII.

OMR (Optical Mark Reader- Nhận dạng dấu quang học) (Krisana Chinnasarn, 1999) là một phương pháp điện tử thu thập dữ liệu do con người xử lý bằng cách xác định một số dấu hiệu nhất định trên tài liệu. Thông thường, quá trình nhận dạng dấu quang học được thực hiện với sự hỗ trợ của máy quét kiểm tra truyền tải hoặc phản xạ ánh sáng qua giấy; những nơi có đánh dấu sẽ phản xạ ít ánh sáng hơn phần giấy trắng, dẫn đến độ tương phản kém hơn. OMR thường được ứng dụng để xử lý dữ liệu từ phiếu điều tra hay chấm các bài thi trắc nghiệm. Ngoài ra, nhiều doanh nghiệp công nghệ cũng đưa ra thuật ngữ Xử lý văn bản thông minh (Intelligent Document Processing-IDP) là một công cụ tự động thu thập, trích xuất dữ liệu từ các tài liệu bán cấu trúc trích xuất nội dung từ ảnh chụp mẫu văn bản có sẵn (giấy chứng minh nhân dân, bằng lái xe, thẻ bảo hiểm y tế, hóa đơn), hoặc theo bất kì định dạng văn bản tùy biến (hợp đồng, chứng từ, quy định...), nhằm số hóa tài liệu một cách nhanh chóng và thuận tiện. FPT.AI Reader được các chuyên gia trong lĩnh vực trí tuệ nhân tạo của FPT Smart Cloud nghiên cứu và phát triển. Giải pháp này áp dụng công nghệ nhận dạng ký tự quang học (OCR) và Xử lý văn bản thông minh (IDP), kết hợp kỹ thuật xử lý ảnh nâng cao và Xử lý ngôn ngữ tự nhiên (NLP), cho phép người dùng số hóa văn bản chính xác trong thời gian ngắn (chỉ tới vài giây) (FPT.AI, 2022). Hình 3 dưới đây là một ví dụ về trích xuất các thông tin cần thiết như họ tên, số căn cước công dân, giới tính, quốc tịch, quê quán, địa chỉ từ ảnh chụp mặt trước của một căn cước công dân theo mẫu hiện tại ở Việt Nam.

Đặc biệt, FPT.AI Reader cho phép người dùng tự định nghĩa mẫu văn bản mới, chưa có trong các mẫu có sẵn của hệ thống để tự tạo mô hình OCR của riêng mình. Chẳng hạn ta

	VIETTEL AI RACE	TD005
	ỨNG DỤNG CÔNG NGHỆ NHẬN DẠNG KÝ TỰ QUANG HỌC	Lần ban hành: 1

cần số hóa thẻ sinh viên, có thể tạo một dự án mới trong bảng điều khiển của FPT.AI (<https://console.fpt.ai/>), tải lên một trúc (semi-structured data) và phi cấu trúc (unstructured data) và chuyển chúng thành tài liệu có cấu trúc (structured data) để sử dụng (Akabot, 2022). IDP là sự kết hợp giữa nhiều công nghệ bao gồm Xử lý ngôn ngữ tự nhiên (Natural Language Processing-NLP), Thị giác máy tính (Computer Vision), Học máy (Machine Learning- ML) và OCR nhằm tăng cường việc nhận diện, phân loại, phân tích, trích xuất dữ liệu và đánh giá dữ liệu để nâng cao độ chính xác và hiệu quả. Ở Việt Nam, nhiều nghiên cứu đã được thực hiện nhằm áp dụng các công nghệ này cho từng loại nghiệp vụ khác nhau, chẳng hạn trích xuất và nhận dạng thông tin trên chứng minh nhân dân của người Việt (Duc Phan, 2021), trích xuất thông tin trên hóa đơn bán hàng (Vu Xuan-Son, 2021), nhận dạng biển số xe (Trần Thị Hương, 2021), trích xuất văn bản từ bìa sách (Phan Thi Thanh Nga, 2017). Tuy nhiên đến nay chưa có nghiên cứu nào áp dụng các công nghệ trên với các tài liệu ở cơ sở giáo dục đại học. Đây cũng là nội dung trọng tâm mà bài báo này hướng đến khi áp dụng thử nghiệm tại Học viện Ngân hàng.

1.3 Kết quả thử nghiệm công nghệ nhận dạng ký tự quang học cho số hóa tài liệu tại Học viện Ngân hàng

1.3.1 Các giải pháp nhận dạng ký tự quang học tại Việt Nam

1.3.1.1. FPT.AI Reader

FPT.AI Reader là ứng dụng nhận dạng và số ảnh mẫu và nhập vào một số trường thông tin quan trọng muốn trích xuất từ ảnh để huấn luyện mô hình OCR (minh họa ở Hình 4).

FPT.AI Reader cũng cho phép tích hợp ứng dụng OCR vào hệ thống của doanh nghiệp thông qua việc đăng ký tài khoản và nhận “API key” từ bảng điều khiển của FPT.AI (Console.fpt.ai). Mỗi API key này cho phép gửi và nhận kết quả đến 50 lần và có thể được mở rộng tùy vào quy mô xử lý dữ liệu của hệ thống.

1.3.1.2. Viettel OCR

Viettel OCR là giải pháp được phát triển bởi bộ phận Trí tuệ nhân tạo của Tập đoàn Viettel, cho phép chuyển đổi tài liệu dạng ảnh (máy quét, máy ảnh, file PDF được chuyển hóa từ file ảnh) thành văn bản như file text (.txt), file Word (.docx). Theo báo cáo trên trang web chính thức, Viettel OCR có khả năng nhận diện văn bản ở dạng bố cục tự do (free layout), có thể tùy chỉnh một cách nhanh chóng để phù hợp với từng bài toán cụ thể của doanh nghiệp. Giải pháp này có thể nhận file đầu vào ở nhiều định dạng khác nhau như PNG, JPEG, cho phép phân tích và trả về kết quả cho nhiều hình ảnh cùng 1 lúc (tối đa 10 ảnh) với độ chính xác tương đối cao (trong báo cáo không ghi rõ độ chính xác).

	VIETTEL AI RACE	TD005
	ỨNG DỤNG CÔNG NGHỆ NHẬN DẠNG KÝ TỰ QUANG HỌC	Lần ban hành: 1

Tuy nhiên trên website chính thức của giải pháp này chưa cho phép người dùng thử nghiệm mà phải liên hệ với tổng đài để đặt lịch tư vấn trực tiếp (Viettel AI, 2021).

1.3.1.3. *Google Vision AI*

Google Vision AI là dịch vụ đám mây của Google cho phép người dùng khởi tạo các ứng dụng phân tích hình ảnh và video trong thời gian ngắn, huấn luyện các mô hình máy học phân loại hình ảnh bằng AutoML hoặc các mô hình tùy chỉnh. Google Vision AI có khả năng phát hiện đối tượng, đọc chữ viết tay và tạo siêu dữ liệu hình ảnh có giá trị bằng các API được đào tạo trước. Một ưu điểm của giải pháp này là dễ dàng tích hợp với BigQuery, Cloud Function và máy ảnh để kích hoạt hành trình từ đầu đến cuối. Google Vision AI là một giải pháp khá linh hoạt, không tạo sẵn các mẫu tài liệu như FPT.AI Reader và Viettel OCR. Tốc độ xử lý của Google Vision AI khá nhanh. Tuy nhiên, quá trình thử nghiệm tính năng OCR của Google Vision AI với một vài tài liệu tiếng Việt cho kết quả chưa tốt (Hình 5).

1.3.1.4. *Lựa chọn giải pháp*

Ngoài 3 giải pháp thương mại kể trên, một số doanh nghiệp công nghệ ở Việt Nam cũng cung cấp các gói dịch vụ số hóa tài liệu với giải pháp riêng mà họ xây dựng. Tuy nhiên qua quá trình tìm hiểu và trải nghiệm, tác giả quyết định lựa chọn FPT. AI Reader làm giải pháp thử nghiệm trong bài báo này với các nguyên nhân:

FPT.AI Reader cung cấp nền tảng sử dụng miễn phí với tất cả người dùng cuối (end-user), chỉ hạn chế số lượng tài liệu xử lý mỗi ngày (50 requests) và có thể mở rộng linh hoạt tùy vào quy mô của doanh nghiệp.

Mô hình OCR huấn luyện sẵn của FPT. AI Reader có khả năng xử lý dữ liệu tiếng Việt tốt (theo báo cáo của FPT.AI đạt trên 96% cho các loại giấy tờ như chứng minh nhân dân, hộ chiếu (FPT.AI, 2022).

Giải pháp FPT.AI Reader đã được rất nhiều khách hàng sử dụng, đa dạng về ngành nghề lĩnh vực như TP Bank, Home Credit, Tiki, EVN, Sendo (FPT.AI, 2022).

1.3.2 **Thử nghiệm với tài liệu tại Học viện Ngân hàng**

1.3.2.1. *Thu thập tài liệu*

Tác giả chọn lọc từ 150 email cá nhân gần nhất có địa chỉ gửi đến từ các đơn vị của Học viện, trích chọn ra những email có tệp đính kèm là các file pdf. Những file này được chọn lọc để đảm bảo đa dạng về nguồn gốc (Học viện ban hành, cơ quan khác ban hành), về

	VIETTEL AI RACE	TD005
	ỨNG DỤNG CÔNG NGHỆ NHẬN DẠNG KÝ TỰ QUANG HỌC	Lần ban hành: 1

định dạng (scan từ máy quét ra hình ảnh, chuyển đổi từ văn bản MS Word), về thể loại (quyết định, phiếu trình, nghị định, thông báo...). Chi tiết số liệu thống kê về bộ tài liệu thử nghiệm được trình bày trong Bảng 1.

1.3.2.2. Cài đặt giải pháp

Giải pháp OCR của FPT.AI được truy cập trên nền tảng điện toán đám mây tại địa [https:// reader.fpt.ai/](https://reader.fpt.ai/). Để tạo ứng dụng mới, FPT.AI Reader cho phép ta chọn trong thư viện những mẫu văn bản có sẵn hoặc tạo ứng dụng tùy chỉnh. Trong thư viện các văn bản có sẵn đã bao gồm:

- Giấy tờ tài chính ngân hàng (như Báo cáo tài chính, Đăng ký kinh doanh...);
- Giấy tờ bảo hiểm và bệnh viện (Giấy ra viện, Phiếu khám, Bảng kê viện phí...);
- Giấy tờ quốc tế (Giấy đăng ký xe, Bằng lái xe của một số nước);
- Giấy tờ tùy thân Việt Nam;
- Giấy tờ khác (Sơ yếu lý lịch, Giấy tờ vận tải, Giấy chứng nhận, Vé máy bay).

Nhận thấy các mẫu văn bản có sẵn trong thư viện không phù hợp với các văn bản trong bộ dữ liệu thử nghiệm, tác giả lựa chọn cài đặt ứng dụng tùy chỉnh. Có 3 mô hình có thể lựa chọn là Bóc tách dữ liệu (từ văn bản có cấu trúc), Mô hình Crop (phát hiện vùng ảnh cần quan tâm) và Phân loại (Gán nhãn phù hợp cho văn bản), tác giả chọn mô hình Bóc tách dữ liệu và sử dụng mô hình Bóc tách dữ liệu OCR có sẵn của FPT.AI chứ không huấn luyện mô hình mới, công việc này sẽ dành cho nghiên cứu trong tương lai với lượng dữ liệu chuẩn bị nhiều và đa dạng hơn.

Trong phần sử dụng mô hình, ta chọn Tải lên để tải văn bản muốn bóc tách, có thể chọn nhiều văn bản cùng lúc, đợi đến khi trạng thái của tất cả văn bản hiện “Thành công” để xem kết quả (minh họa ở Hình 6). Để xem kết quả chi tiết của mỗi tài liệu, ta chọn View ở phần OCR. Một cửa sổ khác sẽ hiện lên hiển thị toàn bộ các trường dữ liệu bóc tách được từ văn bản đầu vào (Hình 7). Trong cửa sổ này ta có thể xem được chi tiết mô hình đã phát hiện được bao nhiêu “box” (vùng ảnh) chứa chuỗi ký tự có thể là văn bản. Click vào chi tiết mỗi box, ta có thể đánh dấu giá trị văn bản phát hiện được trong đó là sai hay đúng, thuộc loại nào (trong trường hợp này chưa định nghĩa loại box nên không hiển thị).

1.3.2.3. Phương pháp đánh giá

Với kết quả nhận được từ ứng dụng vừa xây dựng, tác giả đánh giá bằng cả phương pháp định lượng và định tính theo các tiêu chí sau:

	VIETTEL AI RACE	TD005
	ỨNG DỤNG CÔNG NGHỆ NHẬN DẠNG KÝ TỰ QUANG HỌC	Lần ban hành: 1

Số box phát hiện được có đúng không (có bỏ sót hay phát hiện thừa vùng ảnh nào hay không).

Tỷ lệ sai số theo đơn vị từng từ ở mỗi box (chẳng hạn mô hình dự đoán “Học viện Ngân hàng” so với kết quả đúng là “Học viện Ngân hàng” thì sai số là 25%). Do số lượng từ là quá lớn nên tác giả chỉ thống kê trên trang đầu của mỗi văn bản để kiểm tra những thông tin quan trọng nhất.

Phân tích các trường hợp sai thường gặp theo từng loại văn bản.

1.3.2.4. Kết quả thực nghiệm

Với các tiêu chí kể trên, mô hình của FPT.

AI Reader cho kết quả rất tốt. Cụ thể:

Số box phát hiện được có độ chính xác 98%, không bỏ sót box nào trên các tài liệu. Tuy nhiên một số box bị thừa, đặc biệt là các box ở vùng ảnh chứa con dấu đỏ.

Tỷ lệ sai số khá ấn tượng: 27% với 210 từ bị sai trên tổng số 781 từ (chỉ tính những trang đầu của 20 tài liệu).

Các lỗi sai thường gặp chủ yếu là lỗi dấu câu (“Hoc”- “Học”, “Xã hội”- “Xã hội”...), viết hoa (“đơn VI”- “đơn vị”), và các số quyết định trên phần đầu của văn bản (do là số viết tay).

Các văn bản được chuyển đổi từ file MS Word có tỷ lệ sai số thấp hơn hẳn (trong thực tế những file này có thể chuyển đổi trực tiếp về lại dạng văn bản với các thư viện lập trình phù hợp).

Các trường thông tin quan trọng như đơn vị ban hành văn bản, loại văn bản (thông báo, nghị định, quyết định...), chủ đề phụ của văn bản (về việc điều động, về việc tổ chức...) và đối tượng nhận văn bản có tỷ lệ sai số thấp hơn trung bình chỉ 16%.

Thời gian xử lý của mô hình FPT.AI Reader cũng tương đối nhanh, theo quan sát thực tế, mỗi văn bản 3 trang tốn khoảng trên dưới 5 giây để hoàn thành. Trong thực tế, số trang và dung lượng của văn bản có thể đa dạng và lớn hơn, nhưng để đáp ứng nhu cầu số hóa cơ bản, tức là bóc tách được các trường thông tin quan trọng (thường nằm ở trang đầu văn bản) thì có thể có biện pháp tiền xử lý trước khi đưa vào mô hình.

	VIETTEL AI RACE	TD005
	ỨNG DỤNG CÔNG NGHỆ NHẬN DẠNG KÝ TỰ QUANG HỌC	Lần ban hành: 1

1.4 Kết luận và hướng phát triển

1.4.1 Kết luận

Trong bài báo này, tác giả đã nêu tầm quan trọng của việc số hóa dữ liệu, khảo sát sơ bộ thực trạng số hóa tài liệu văn bản tại Học viện Ngân hàng. Qua đó tác giả đặt ra 4 mục tiêu nghiên cứu và đã giải quyết lần lượt từng mục tiêu. Có thể tóm tắt một số đóng góp chính của bài báo như sau:

Giới thiệu các công nghệ cốt lõi như nhận dạng ký tự quang học (OCR), nhận dạng dấu quang học (OMR), xử lý văn bản thông minh (IDP) và những ứng dụng trên nhiều lĩnh vực của các công nghệ này.

Tìm hiểu và khảo sát một số giải pháp thương mại cho số hóa tài liệu phổ biến ở Việt Nam như FPT.AI Reader, ViettelOCR, Google Vision AI. Trên cơ sở đó tác giả so sánh đánh giá và lựa chọn FPT.AI Reader để cài đặt ứng dụng thử nghiệm tại Học viện Ngân hàng.

Thu thập dữ liệu văn bản từ 150 email nội bộ của Học viện, chọn lọc 20 file pdf đính kèm trong các email đó sao cho đảm bảo có các loại văn bản khác nhau, các định dạng khác nhau và các phòng ban phụ trách khác nhau.

Cài đặt thử nghiệm hệ thống OCR trên nền tảng FPT.AI Reader để áp dụng với bộ dữ liệu văn bản vừa thu thập. Kết quả thử nghiệm cho thấy mô hình nhận dạng ký tự của FPT.AI hoạt động khá tốt với 27% sai số ở mức từ và thời gian xử lý chỉ khoảng 5 giây cho mỗi văn bản 3 trang, chỉ tồn tại một vài sai số ở những ký tự viết tay hoặc bị mờ do chất lượng scan tài liệu.

1.4.2 Hướng phát triển

Với kết quả thử nghiệm trình bày như trên, tác giả đề xuất Học viện Ngân hàng đẩy mạnh áp dụng các công nghệ mới cho quá trình số hóa dữ liệu, đặc biệt là dữ liệu văn bản. Giải pháp của FPT.AI tuy còn một số tồn tại, sai số, nhưng có thể được cải thiện bằng cách huấn luyện hoặc tinh chỉnh mô hình OCR với bộ dữ liệu đầy đủ và đa dạng hơn của Học viện. Điều này sẽ cần các nghiên cứu chuyên sâu về công nghệ OCR trong tương lai. ■

2. TÀI LIỆU THAM KHẢO

Akabet (2022), Sự khác biệt giữa OCR và IDP, Truy cập ngày 20 tháng 4 năm 2023, từ <https://akabet.com/vi/tai-nguyen/blog/su-khac-biet-giua-ocr-va-idp/>

	VIETTEL AI RACE	TD005
	ỨNG DỤNG CÔNG NGHỆ NHẬN DẠNG KÝ TỰ QUANG HỌC	Lần ban hành: 1

ALS (2021), Quy trình các bước số hóa tài liệu lưu trữ doanh nghiệp, Truy cập ngày 20 tháng 4 năm 2023, từ <https://als.com.vn/quy-trinh-cac-buoc-so-hoa-tai-lieu-luu-tru-doanh-nghiep>

Duc Phan Van Hoai, Huu-Thanh Duong, Vinh Truong Hoang (2021), “Text recognition for Vietnamese identity card based on deep features network”, International Journal on Document Analysis and Recognition (IJDAR), số 24, trang 123–131.

FPT.AI (2022), FPT.AI Reader - Vietnamese Passport Recognition, Truy cập ngày 20 tháng 4 năm 2023, từ <https://docs.fpt.ai/docs/en/vision/documentation/license-recognition>

FPT.AI (2022), “FPT AI Read - Trích xuất dữ liệu vượt mọi giới hạn”, Truy cập ngày 20 tháng 4 năm 2023, từ <https://fpt.ai/vi/reader>

FPT.AI (2022), Hướng dẫn sử dụng FPT.AI Reader - phần mềm ocr trích xuất thông tin từ ảnh chụp, Truy cập ngày 20 tháng 4 năm 2023, từ <https://fpt.ai/vi/huong-dan-su-dung-fptai-reader-phan-mem-ocr-trich-xuat-thong-tin-tu-anh-chup>

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, Seunghyun Park (2022), “OCR-Free Document Understanding Transformer”, Computer Vision – ECCV, số 13688, trang 498–517.

Học viện Ngân hàng (2023), Tờ trình số 694/TTr-HVNH ngày 28/3/2023 Kế hoạch Chuyển đổi số tại Học viện Ngân hàng. Krisana Chinnasarn, Yuttapong Rangsanteri (1999), “Image-processing-oriented optical mark reader”. Applications of Digital Image Processing XXII, số 3808.

Noman Islam, Zeeshan Islam, Nazia Noor (2016), “A Survey on Optical Character Recognition System”, Journal of Information & Communication Technology-JICT, số 10, trang 1-4

Phạm Huy Giao (2020), “Chuyển đổi số: Bản chất, thực tiễn và ứng dụng”, Tạp chí Dầu khí, số 12, trang 12-16. Phan Thi Thanh Nga, Nguyễn Thị Huyền Trang, Nguyễn Văn Phúc, Thái Duy Quý, Võ Phương Bình (2017), “Vietnamese

text extraction from book covers”. Tạp chí Khoa học Đại học Đà Lạt”, số 7, trang 142–152.

	VIETTEL AI RACE	TD005
	ỨNG DỤNG CÔNG NGHỆ NHẬN DẠNG KÝ TỰ QUANG HỌC	Lần ban hành: 1

Ravina Mithe, Supriya Indalkar, Nilam Divekar (2013), “Optical Character Recognition”, International Journal of Recent Technology and Engineering (IJRTE), số 2, trang 72-75.

Raymond Ptucha, Felipe Petroski Such, Suhas Pillai, Frank Brockler, Vatsala Singh, Paul Hutkowsky (2019), “Intelligent character recognition using fully convolutional neural networks”, Pattern Recognition, số 88, trang 604-613.

Thủ tướng Chính phủ (2022), Quyết định số 131/QĐ-TTg ngày 25/01/2022 của Thủ tướng Chính phủ: Phê duyệt Đề án “Tăng cường ứng dụng công nghệ thông tin và chuyển đổi số trong giáo dục và đào tạo giai đoạn 2022-2025, định hướng đến năm 2030”

Trần Thị Hương, Ngô Thị Kiều Hằng (2021), “Kỹ thuật nhận dạng biển số xe và ứng dụng vào bài toán quản lý bãi giữ xe tại trường đại học Hà Tĩnh”. Tạp chí Khoa học Đại học Đồng Tháp, số 3, trang 115-120.

Viettel AI (2021), “Nhận dạng ký tự quang học”, Truy cập ngày 20 tháng 4 năm 2023, từ <https://viettelgroup.ai/service/ocr>

Vu Xuan-Son, Bui Quang-Anh, Nguyen Nhu-Van, Hai Nguyen Thi Tuyet, Vu Thanh (2021), “MC-OCR Challenge:

Mobile-Captured Image Document Recognition for Vietnamese Receipts”, RIVF International Conference on Computing and Communication Technologies, IEEE, trang 1-6.