Đồ án tốt nghiệp

Duc Ho V

July 2023

MỤC LỤC

MỞ ĐẦU

CHƯƠNG 1: GIỚI THIỆU

1.1 Lý do chọn đề tài

Cuộc sống hiện nay việc mua bán trao đổi hàng hóa được diễn ra thường xuyên giữa người mua và người bán. Ban đầu hóa đơn có giá trị làm bằng chứng chứng nhận cho việc chuyển nhượng hàng hóa giữa hai bên, có giá trị làm bằng chứng chứng nhận cho việc chuyển nhượng hàng hoá giữa hai bên. Mọi việc tranh chấp trong mua bán hàng hoá hai bên tự giải quyết.

Trong quá trình phát triển xã hội, hoá đơn được phổ biến dần trong một cộng đồng khi được cộng đồng chấp nhận một cách tự nguyện. Các cộng đồng có thể là các Phường hội hoặc các định chế làng, xã. Những tranh chấp trong việc mua bán hàng hoá được các cộng đồng xử lý trên cơ sở dân sự. Khi nhà nước tham dự vào quản lý mua bán hàng hoá và xử lý những tranh chấp về hàng hoá dựa trên pháp luật dân sự và hình sự thì hoá đơn được nhà nước quy định để làm căn cứ pháp lý chứng minh cho việc chuyển nhượng hàng hoá giữa các bên và làm căn cứ để xác nhận quyền sở hữu hợp pháp của người có hàng hoá. Do đó hóa đơn là một loại tài liệu quan trọng trong các giao dịch. Nó được sử dụng để ghi lại các giao dịch mua bán hàng hóa và dịch vụ. Thông tin trên hóa đơn bao gồm tên của người bán, tên của người mua, ngày lập hóa đơn, số lượng hàng hóa hoặc dịch vụ, giá cả, tổng số tiền phải thanh toán.v.v...

Hiện nay, hóa đơn thông thường được lập dưới dạng tài liệu giấy, có thể là hóa đơn giá trị gia tăng, hóa đơn bán hàng, tem, vé, thẻ, phiếu thu tiền bảo hiểm... Hóa đơn giấy có thể được phát hành theo các hình thức như hóa đơn đặt in, hóa đơn tự in, hóa đơn mua của cơ quan thuế. Điều này gây ra một số khó khăn trong việc quản lý hóa đơn, chẳng hạn như:

- Quá nhiều hóa đơn: Các doanh nghiệp có thể phát sinh một số lượng lớn hóa đơn, từ các nhà cung cấp, khách hàng và các bên liên quan khác. Việc quản lý nhiều hóa đơn có thể là một thách thức, đặc biệt nếu chúng không được tổ chức và lưu trữ một cách hiệu quả.
- Sai sót và mất mát: Quản lý hóa đơn thủ công có thể gặp phải sai sót và mất mát hóa đơn, đặc biệt khi các hóa đơn được lưu trữ và xử lý bằng tay. Điều này có thể dẫn đến việc đòi tiền sai, không thu được tiền đúng lúc hoặc mất cơ hội thu hồi tiền nợ.
- Tìm kiếm hóa đơn: Khi cần tìm một hóa đơn cụ thể, việc tìm kiếm nó có thể là một thách thức nếu nó không được tổ chức và lưu trữ một cách hiệu quả. Điều này có thể dẫn đến chậm trễ trong quá trình thanh toán hóa đơn hoặc thậm chí mất hóa đơn.
- Lưu trữ hóa đơn. Các hóa đơn phải được lưu trữ trong một thời gian nhất định theo quy định của pháp luật. Điều này có thể là một thách thức nếu không có một quy trình lưu trữ hóa đơn hiệu quả.

- Tuân thủ luật pháp: Việc tuân thủ các quy định và luật pháp về hóa đơn là rất quan trọng. Nếu không tuân thủ đúng, doanh nghiệp có thể phải đối mặt với các vấn đề pháp lý và hậu quả tài chính nghiêm trọng.
- Thay đổi trong quy định thuế: Thay đổi trong quy định thuế và các quy tắc về hóa đơn có thể làm cho việc quản lý hóa đơn trở nên phức tạp hơn, đòi hỏi doanh nghiệp phải cập nhật và điều chỉnh quy trình của mình thường xuyên.

Để giải quyết các vấn đề này ta có thể cân nhắc sử dụng các phần mềm quản lý hóa đơn hiện đại. OCR có thể giải quyết vấn đề này bằng cách tự động hóa quá trình nhập liệu của hóa đơn một cách đơn giản và dễ dàng.

1.2 Giải pháp

OCR là một công nghệ có thể giải quyết các vấn đề trên bằng cách tự động trích xuất thông tin từ hóa đơn. Đây là một công nghệ cho phép máy tính nhận dạng và chuyển đổi văn bản từ hình ảnh chứa văn bản thành dạng văn bản có thể chỉnh sửa, tìm kiếm và lưu trữ. Áp dụng OCR trong việc quản lý hóa đơn có thể giúp giải quyết một số vấn đề như sau:

- **Tiết kiệm thời gian**: OCR giúp doanh nghiệp tiết kiệm lượng lớn thời gian so với quá trình nhập dữ liệu thủ công. Với công cụ OCR, thông tin có thể dễ dàng được trích xuất sang các định dạng kỹ thuật số theo nhu cầu chỉ bằng việc chụp và tải ảnh lên. Không chỉ vậy, dữ liệu khi được trích xuất có thể dễ dàng được tìm kiếm, chỉnh sửa và thực hiện nhiều tác vụ khác, hỗ trợ quy trình xử lý tài liệu dễ dàng và thuận tiện hơn. Trên thực tế, nghiên cứu đã phát hiện ra rằng lượng thời gian dành cho công việc giấy tờ có thể giảm 75% khi sử dụng OCR. Trung bình, thời gian để trích xuất một tài liệu sang dạng số chỉ từ 0.5 – 2 giây với công cụ OCR, một sự tối ưu đáng kể so với thời giang trung bình 1– 5 phút khi sử dụng phương pháp nhập liệu truyền thống. [1]
- **Cải thiện độ chính xác**: Việc nhập liệu bằng tay không chỉ tốn nhiều thời gian, nguồn lực mà còn có mức độ rủi ro cao trong sai sót nhập. Nhất là với các loại tài liệu bao gồm nhiều trường thông tin bằng số, địa chỉ email, địa chỉ nhà,... việc nhập tay thủ công khó có thể chính xác 100%. Những lỗi sai thông tin ngay từ bước đầu sẽ khiến kho dữ liệu doanh nghiệp không được "sạch" và chính xác.
- **Hỗ trợ tuân thủ luật pháp**: Sử dụng OCR giúp đảm bảo tính chính xác và toàn vẹn của dữ liệu trên hóa đơn, từ đó đảm bảo tuân thủ các quy định về hóa đơn và thuế.
- **Quản lý hóa đơn điện tử**: Kết hợp OCR với hóa đơn điện tử giúp tự động tạo và lưu trữ các hóa đơn điện tử, giảm thiểu việc sử dụng giấy tờ truyền thống và tiết kiệm không gian lưu trữ.

Nhìn chung, OCR là một công nghệ có nhiều tiềm năng ứng dụng trong lĩnh vực kế toán và tài chính. OCR có thể giúp các doanh nghiệp tiết kiệm thời gian, tăng cường độ chính xác và cải thiện khả năng truy xuất thông tin hóa đơn. Tuy nhiên, để OCR hiệu quả, ta cần đảm bảo rằng hóa đơn được quét và lưu trữ ở định dạng tốt, đủ để đảm bảo hiệu suất nhận dạng của OCR cao nhất.

1.3 Mục Tiêu

Dựa vào những vấn đề của hóa đơn và các giải pháp của OCR ở mục 1.1 và 1.2, mục tiêu của đề tài "**Nghiên cứu ứng dụng công nghệ OCR nhận dạng hóa đơn**" là tìm hiểu, đánh giá khả năng ứng dụng của công nghệ OCR hiện nay trong việc quản lý hóa đơn. Cụ thể đề tài tập trung vào các mục tiêu sau:

- Tìm hiểu về công nghệ OCR: Nghiên cứu các nguyên lý hoạt động của OCR, các phương pháp và thuật toán phổ biến trong việc nhận dạng văn bản từ hình ảnh.
- Phân tích hiệu quả và lợi ích của ứng dụng OCR trong quản lý hóa đơn: So sánh các phương pháp truyền thống và ứng dụng OCR trong việc quản lý hóa đơn, đánh giá hiệu quả và lợi ích mà OCR mang lại, bao gồm tối ưu hóa thời gian, giảm thiểu sai sót, tiết kiệm chi phí và tăng cường khả năng xử lý lượng hóa đơn lớn.
- Đề xuất giải pháp và quy trình triển khai OCR: Dựa trên kết quả nghiên cứu, đề xuất các giải pháp và quy trình triển khai OCR trong việc quản lý hóa đơn, bao gồm lựa chọn phần mềm OCR phù hợp, quy trình xử lý hóa đơn, quản lý dữ liệu và bảo đảm tính an toàn thông tin.
- Đánh giá hiệu quả thực tế: Tiến hành thử nghiệm ứng dụng OCR trong môi trường thực tế của doanh nghiệp hoặc tổ chức để đánh giá hiệu quả, tính ổn định và khả năng mở rộng của giải pháp OCR.

Dựa trên kết quả đánh giá, đề xuất các cải tiến và phát triển tương lai của công nghệ OCR trong việc quản lý hóa đơn, nhằm nâng cao hiệu quả và khả năng ứng dụng của nó trong thực tế

1.4 Phạm vi đề tài

1.5 Bố cục chương

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

Với sự phát triển không ngừng của công nghệ, việc chuyển đổi thông tin từ tài liệu giấy trở thành dạng điện tử đã trở nên cực kỳ quan trọng đối với doanh nghiệp và cá nhân. Trước đây, việc nhập liệu thủ công từ hóa đơn và tài liệu tương tự tốn rất nhiều thời gian, công sức và có nguy cơ sai sót cao. Tuy nhiên, với sự xuất hiện của công nghệ OCR, quá trình này đã trở nên tự động và hiệu quả hơn, giúp tiết kiệm thời gian và tối ưu hóa quá trình làm việc.

Trong chương này, ta sẽ bắt đầu bằng việc tìm hiểu về cơ bản của công nghệ OCR, một công nghệ quan trọng đã thay đổi cách chúng ta xử lý và quản lý thông tin, khám phá cách OCR có thể phân tích hình ảnh, xác định kí tự và từ, và biến đổi chúng thành dạng văn bản có thể xử lý.

Qua đó, chương cơ sở lý thuyết này sẽ tạo nền tảng kiến thức quan trọng cho việc hiểu rõ hơn về công nghệ OCR và tầm quan trọng của nó trong việc tối ưu hóa quá trình nhận dạng hóa đơn.

2.1 Nhận dạng ký tự quang học

2.1.1 Nhận dạng ký tự quang học là gì?

Nhận dạng ký tự quang học hay còn gọi là OCR đây là quá trình chuyển đổi một hình ảnh văn bản viết tay, đánh máy hoặc in thành định dạng văn bản mà máy có thể hiểu được. Nó được sử dụng rộng rãi để nhận dạng và tìm kiếm văn bản từ các tài liệu điện tử hoặc để xuất bản văn bản trên một trang web. [2], [3]

OCR được sử dụng rộng rãi như một hình thức nhập dữ liệu từ các bản ghi dữ liệu giấy in - cho dù đó là tài liệu hộ chiếu, hóa đơn, sao kê ngân hàng, biên lai vi tính hóa, danh thiếp, thư, dữ liệu in hoặc bất kỳ tài liệu phù hợp nào - đó là một phương pháp phổ biến để số hóa các văn bản in sao cho chúng có thể được chỉnh sửa, tìm kiếm, lưu trữ bằng điện tử, hiển thị trực tuyến và được sử dụng trong các quy trình máy như điện toán nhận thức, dịch máy, chuyển văn bản thành giọng nói (trích xuất), dữ liệu chính và khai thác văn bản. OCR là một lĩnh vực nghiên cứu về nhận dạng mẫu, trí tuệ nhân tạo và thị giác máy tính.[4]

Nhận dạng ký tự quang học đã được áp dụng vào nhiều ứng dụng khác nhau. Dưới đây là một số ứng dụng của OCR: [3]

- **Nhận dạng chữ viết tay:** Máy tính để nhận và diễn dịch thông tin viết tay rõ ràng từ các nguồn như tài liệu giấy, ảnh, màn hình cảm ứng và các thiết bị khác. Hình ảnh văn bản viết có thể được cảm nhận "ngoại tuyến" từ tờ giấy thông qua quét quang học hoặc nhận dạng từ thông minh. Một cách khác, các chuyển động của đầu bút viết có thể được cảm

nhận "trực tuyến", ví dụ như bề mặt màn hình máy tính dựa trên bút viết.

- **Ngân hàng:** Được sử dụng để xử lý séc mà không cần sự tham gia của con người. Một tờ séc có thể được đặt vào máy, trong đó hệ thống quét số tiền cần phát hành và số tiền chính xác sẽ được chuyển khoản. Công nghệ này đã gần như được hoàn thiện cho các séc được in ấn và cũng khá chính xác đối với các séc viết tay, giảm thiểu thời gian chờ đợi tại ngân hàng.
- **Chăm sóc sức khỏe:** Các chuyên gia y tế luôn phải đối mặt với số lượng lớn các biểu mẫu cho mỗi bệnh nhân, bao gồm cả biểu mẫu bảo hiểm cũng như các biểu mẫu sức khỏe chung. Để theo kịp với tất cả thông tin này, việc nhập dữ liệu liên quan vào một cơ sở dữ liệu điện tử có thể được truy cập khi cần thiết. Các công cụ xử lý biểu mẫu, được cung cấp bởi công nghệ OCR, có khả năng trích xuất thông tin từ các biểu mẫu và đưa vào cơ sở dữ liệu, để mỗi dữ liệu bệnh nhân được ghi lại đúng thời điểm.
- **Captcha:** Trong CAPTCHA, một hình ảnh gồm các ký tự hoặc số được tạo ra, bị mờ đi bằng các kỹ thuật biến dạng hình ảnh, biến đổi kích thước và phông chữ, phông nền gây xao lãng, đoạn ngẫu nhiên, đánh dấu và nhiễu trong hình ảnh. Hệ thống này có thể được sử dụng để loại bỏ nhiễu và phân đoạn hình ảnh để làm cho hình ảnh dễ xử lý cho các hệ thống OCR
- **Ảnh hóa đơn:** Được sử dụng rộng rãi trong nhiều ứng dụng kinh doanh để theo dõi hồ sơ tài chính và ngăn chặn việc tích lũy các khoản thanh toán chồng chất.
- **Nhận dạng biển số xe**: Sử dụng để tự động nhận dạng và ghi nhận biển số xe trên các hình ảnh hoặc video.
- ...

Từ những ứng dụng trên ta có thể thấy rằng OCR đang được sử dụng rộng rãi trong cuộc sống hàng ngày, nó đang đóng vai trò quan trọng trong việc chuyển đổi số hiện nay. Điều này rất quan trọng để tối ưu hóa quá trình làm việc với thông tin trong thời đại công nghệ thông tin.

2.1.2 Lịch sử của OCR

OCR được ra đời và cuối thế kỉ 19, được cấp bằng sáng chế tại Mỹ vào ngày 31 tháng 12 năm 1935 của Gustav Tauschek đến từ Viên, Áo, đây là một trong những phát minh sớm nhất liên quan đến OCR. OCR ban đầu được sử dụng để số hóa các văn bản in và cho phép chúng có thể đọc được bằng máy. Khi công nghệ OCR tiếp tục phát triển, nó đã được sử dụng rộng rãi trong các ngành công nghiệp khác nhau.

Sự khởi đầu thực sự của những hệ thống OCR ban đầu thực sự bắt đầu vào những năm 1960 và 1970. Các hệ thống này được thiết kế cho các trường hợp sử dụng cụ thể, chẳng hạn như phân loại thư dựa trên mã zip hoặc đọc số viết tay. Phông chữ có thể đọc bằng máy quang học đầu tiên OCR-A được phát triển vào năm 1968 bởi nhà thiết kế kiểu chữ người Thụy Sĩ Adrian Frutiger.

Trong suốt những năm 1980, công nghệ OCR đã đạt được những bước tiến đáng kể với sự phát triển của các thuật toán mới và các máy tính mạnh hơn. Các hệ thống OCR có thể nhận dạng nhiều loại phông chữ hơn và có thể xử lý các hình ảnh phức tạp hơn, khiến chúng trở nên chính xác và hữu ích hơn cho nhiều ứng dụng hơn.

Vào những năm 1990, việc sử dụng rộng rãi máy tính cá nhân và internet đã dẫn đến sự gia tăng đáng kể trong việc sử dụng công nghệ OCR. Các hệ thống OCR được sử dụng để số hóa sách, tạp chí và các tài liệu in khác, giúp tìm kiếm và truy cập thông tin dễ dàng hơn. Công nghệ này cũng được sử dụng để tự động hóa các quy trình nhập dữ liệu trong các ngành như tài chính, chăm sóc sức khỏe và chính phủ.

Vào đầu những năm 2000, lịch sử của công nghệ OCR đã phát triển với việc giới thiệu các thuật toán mới và phần cứng được cải tiến. Các hệ thống OCR trở nên chính xác hơn và có thể nhận dạng nhiều loại ký tự và ngôn ngữ hơn. Điều này đã mở đường cho việc áp dụng rộng rãi công nghệ OCR trong nhiều ngành và ứng dụng khác nhau, chẳng hạn như quản lý tài liệu và xử lý hóa đơn. Trong khung thời gian này, Google cũng nổi tiếng (và gây tranh cãi) đã ra mắt Google Sách, có tên mã là Dự án Đại dương, sử dụng OCR để số hóa hàng chục triệu cuốn sách và làm cho văn bản của chúng có thể tìm kiếm được.

Ngày nay, công nghệ OCR tiên tiến và phức tạp hơn bao giờ hết. Các hệ thống OCR có thể nhận dạng nhiều loại ký tự và ngôn ngữ, chữ viết tay và các hình ảnh phức tạp khác. Công nghệ OCR đang tiếp tục phát triển và những tiến bộ mới nhất về trí tuệ nhân tạo và máy học đang dẫn đến các hệ thống thậm chí còn phức tạp và chính xác hơn.

Lịch sử OCR bắt đầu với những phát minh mang tính cách mạng được thiết kế để cải thiện chất lượng cuộc sống cho nhân loại. Nhiều thập kỷ sau, công nghệ này vẫn đang trải qua quá trình phát triển và cải tiến liên tục, đồng thời là một yếu tố quyết định quan trọng của thời đại kỹ thuật số. OCR đã trải qua một chặng đường dài và đang thực sự cải thiện chất lượng cuộc sống của phần lớn nhân loại. Ngày nay, nhiều ngành công nghiệp và ứng dụng sử dụng OCR. Trong những thập kỷ tới, nó sẽ đóng một vai trò quan trọng trong quá trình chuyển đổi kỹ thuật số toàn cầu.[5]

2.1.3 Nguyên tắc hoạt động của OCR

OCR hoạt động bằng cách phân tích hình ảnh văn bản và sau đó tạo ra một bản sao văn bản kỹ thuật số của hình ảnh đó. Quá trình này thường

được thực hiện theo các bước sau:

- Quét tài liệu: Tài liệu được quét bằng máy quét để tạo ra một hình ảnh kỹ thuật số của tài liệu.
- Phân tích và xử lý hình ảnh: Trước khi nhận dạng văn bản, ảnh được tiền xử lý để làm sạch và cải thiện chất lượng. Điều này có thể bao gồm việc điều chỉnh độ tương phản, loại bỏ nhiễu, cắt biên và xoay ảnh để đảm bảo văn bản nằm ngang. Sau đó hình ảnh được phân tích để xác định các vùng văn bản.
- Nhận dạng ký tự: Trong bước này, hình ảnh được chuyển đổi thành dạng dữ liệu văn bản bằng cách nhận dạng các ký tự riêng lẻ. Các thuật toán và mô hình máy học được sử dụng để so khớp các đặc trưng trong hình ảnh với các ký tự đã biết từ bộ dữ liệu huấn luyện.
- Phân tích cấu trúc: Sau khi xác định được các ký tự, công cụ OCR cũng cố gắng xác định cấu trúc của văn bản, bao gồm việc xác định các đoạn, đoạn văn bản, tiêu đề, danh sách và các yếu tố cấu trúc khác.
- Sửa lỗi và kiểm tra: Sau khi nhận dạng, dữ liệu văn bản thường cần được kiểm tra lại và sửa lỗi do các lỗi nhận dạng có thể xảy ra. Điều này có thể thực hiện tự động hoặc thông qua giao diện người dùng để đảm bảo tính chính xác của kết quả.
- Tạo bản sao văn bản kỹ thuật số: Một bản sao văn bản kỹ thuật số của hình ảnh được tạo ra bằng cách kết hợp các ký tự đã được nhận dạng.

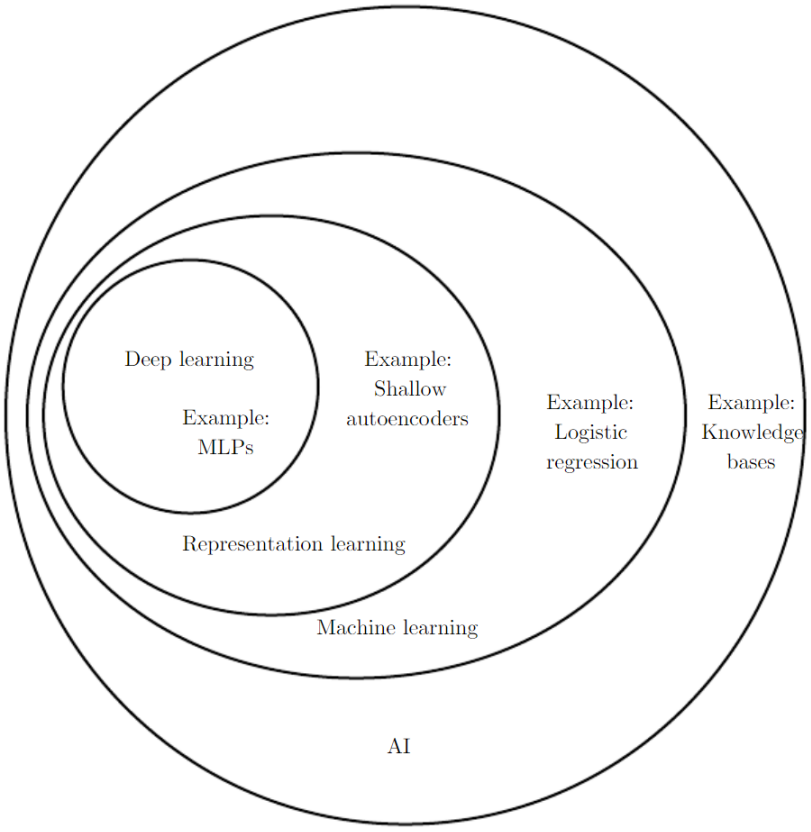
Các công nghệ OCR ngày càng phát triển, sử dụng các mô hình học sâu và học máy để cải thiện khả năng nhận dạng và xử lý ngôn ngữ tự nhiên, tạo ra kết quả chính xác hơn và phức tạp hơn.

2.2 Học sâu và ứng dụng trong OCR

Trong việc ứng dụng OCR để nhận dạng hóa đơn, mạng học sâu đã chơi một vai trò quan trọng và mang lại những cải tiến đáng kể cho quá trình này. Trước khi sự xuất hiện của học sâu, các hệ thống nhận dạng dựa trên các phương pháp truyền thống thường gặp khó khăn trong việc xử lý các biến thể phức tạp của hình ảnh hóa đơn và khả năng xử lý đa dạng của chúng. Nhưng với mạng nơron học sâu, khả năng học và tự điều chỉnh của mô hình đã mở ra những cánh cửa mới cho việc nhận dạng hóa đơn hiệu quả hơn.

2.2.1 Học sâu là gì?

Học sâu là một các tiếp cận của Trí tuệ nhân tạo. Cụ thể thì nó là một kiểu của học máy (Hình 1), một kỹ thuật mà cho phép hệ thống máy tính tự học từ trải nghiệm và dữ liệu, nó sở hữu sức mạnh và sự linh hoạt tuyệt vời thông qua việc học cách biểu diễn như một hệ phân cấp khái niệm trong



Hình 1: Biểu đồ Venn về Trí tuệ nhân tạo

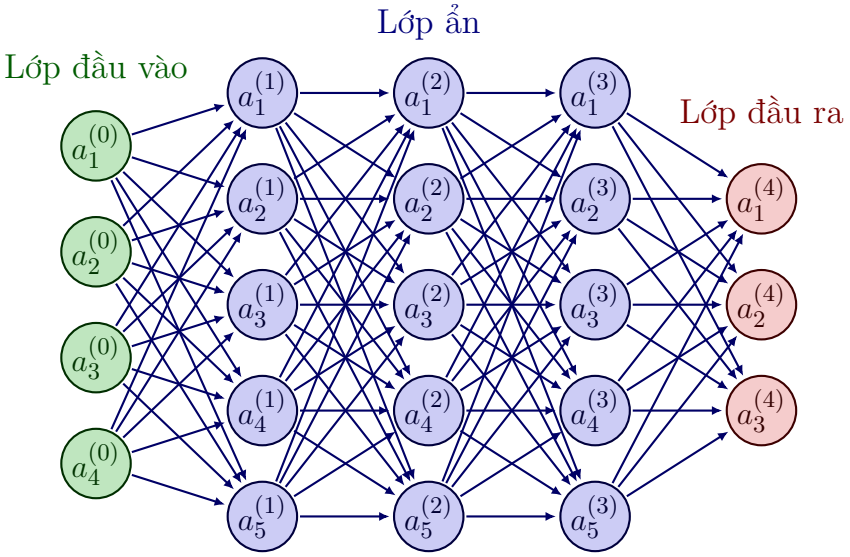
đó mỗi khái niệm được định nghĩa từ những khái niệm đơn giản hơn, và mỗi biểu diễn được tính tính toán từ những biểu diễn kém trừu tượng hơn. [6]

Một mạng học sâu bao gồm những thành phần sau đây [7]:

- **Lớp đầu vào:** Một mạng nơ-ron nhân tạo sẽ có một số nút để nhập dữ liệu đầu vào. Các nút này tạo nên lớp đầu vào của hệ thống.
- **Lớp ẩn:** Lớp đầu vào xử lý và chuyển dữ liệu đến các lớp sâu hơn trong mạng nơ-ron. Các lớp ẩn này xử lý thông tin ở các cấp độ khác nhau, thích ứng với hành vi của mình khi nhận được thông tin mới. Các mạng học sâu có hàng trăm lớp ẩn có thể được dùng để phân tích một vấn đề từ nhiều góc độ khác nhau.
- **Lớp đầu ra:** Lớp đầu ra bao gồm các nút xuất dữ liệu. Các mô hình học sâu xuất ra đáp án "có" hoặc "không" chỉ có hai nút trong lớp đầu ra. Mặt khác, các mô hình xuất ra nhiều đáp án hơn sẽ có nhiều nút hơn.

Một điểm đáng chú ý là học sâu cần một lượng lớn dữ liệu để huấn luyện mô hình một cách hiệu quả [8]. Trong trường hợp OCR và nhận dạng hóa đơn, mạng nơron học sâu có khả năng học từ hàng nghìn hoặc thậm chí hàng triệu hình ảnh hóa đơn, điều này giúp mô hình hiểu rõ các đặc trưng và biểu diễn của dữ liệu hơn.

Hiện nay các kiến trúc học sâu như mạng nơ-ron sâu, mạng niềm tin sâu,



Hình 2: Kiến trúc một mạng Nơ-ron

học tăng cường sâu, mạng nơ-ron tái phát, mạng nơ-ron tích chập và máy biến áp đã được áp dụng cho các lĩnh vực bao gồm thị giác máy tính, nhận dạng giọng nói, xử lý ngôn ngữ tự nhiên, dịch máy, tin sinh học, thiết kế thuốc, Phân tích hình ảnh y tế, khoa học khí hậu, kiểm tra vật liệu và các chương trình trò chơi trên bàn cờ, nơi chúng đã tạo ra kết quả tương đương và trong một số trường hợp vượt qua hiệu suất chuyên gia của con người.

2.2.2 Ứng dụng trong OCR

Học sâu đã được chứng minh là có hiệu quả hơn các phương pháp OCR truyền thống [9], đặc biệt là đối với các tài liệu có chất lượng thấp hoặc bị định dạng phức tạp.

Một số ứng dụng của Học sâu trong OCR bao gồm:

- Nhận dạng hóa đơn: Học sâu có thể được sử dụng để nhận dạng các trường thông tin quan trọng trên hóa đơn, chẳng hạn như tên người mua, người bán, ngày giao dịch, số lượng, giá và tổng số tiền. Điều này có thể giúp tiết kiệm thời gian và chi phí cho các doanh nghiệp, đồng thời cải thiện độ chính xác của việc xử lý hóa đơn.
- Nhận dạng tài liệu y tế: Học sâu có thể được sử dụng để nhận dạng thông tin quan trọng trên tài liệu y tế, chẳng hạn như tên bệnh nhân, chẩn đoán, phương pháp điều trị và các loại thuốc được kê đơn. Điều này có thể giúp cải thiện chất lượng chăm sóc bệnh nhân và giảm chi phí chăm sóc sức khỏe.
- Nhận dạng tài liệu pháp lý: Học sâu có thể được sử dụng để nhận dạng thông tin quan trọng trên tài liệu pháp lý, chẳng hạn như tên các bên liên quan, ngày tháng, các điều khoản của thỏa thuận và các điều khoản

của hợp đồng. Điều này có thể giúp các luật sư và chuyên gia pháp lý tìm kiếm thông tin nhanh chóng và dễ dàng hơn.

- Nhận dạng tài liệu tài chính: Học sâu có thể được sử dụng để nhận dạng thông tin quan trọng trên tài liệu tài chính, chẳng hạn như tên công ty, giá cổ phiếu, số lượng cổ phiếu và giá trị thị trường của cổ phiếu. Điều này có thể giúp các nhà đầu tư và các chuyên gia tài chính đưa ra quyết định đầu tư tốt hơn.

Học sâu là một công nghệ mạnh mẽ có thể được sử dụng để cải thiện độ chính xác và hiệu quả của OCR. Với sự phát triển của Học sâu, OCR sẽ trở nên dễ dàng và thuận tiện hơn trong tương lai.

2.3 Các thuật toán OCR

Mặc dù OCR tương đối cụ thể, nhưng nó liên quan đến nhiều khía cạnh của công nghệ, bao gồm phát hiện văn bản, nhận dạng văn bản, nhận dạng văn bản từ đầu đến cuối, phân tích tài liệu, v.v. Nghiên cứu học thuật về các công nghệ liên quan của OCR phát triển mạnh mẽ. Phần này đây sẽ giới thiệu sơ lược về một số công nghệ chính trong tác vụ OCR.

2.3.1 Phát hiện văn bản

Công việc phát hiện văn bản là để xác định vùng chứa văn bản trên ảnh đầu vào. Trong những năm gần đây, có nhiều nghiên cứu học thuật về phát hiện văn bản. Một lớp phương pháp coi việc phát hiện văn bản như một tình huống cụ thể trong việc phát hiện mục tiêu, và điều chỉnh các thuật toán phát hiện mục tiêu chung để phù hợp với việc phát hiện văn bản. Ví dụ, TextBoxes dựa trên một bộ phát hiện mục tiêu một giai đoạn là SSD. Thuật toán điều chỉnh khung mục tiêu để vừa với các dòng văn bản có tỷ lệ khía cạnh cực đoan, trong khi CTPN được phát triển từ Faster RCNN. Tuy nhiên, vẫn có một số khác biệt giữa phát hiện văn bản và phát hiện mục tiêu về thông tin mục tiêu và nhiệm vụ chính. Ví dụ, văn bản thường dài và trông giống "vạch", khoảng cách giữa các dòng nhỏ, văn bản có thể uốn cong, v.v. Do đó, nhiều thuật toán đặc biệt cho việc phát hiện văn bản đã được phát triển, như EAST, PSENet, DBNet, và nhiều thuật toán khác [10].

Hiện tại, một số thuật toán phát hiện văn bản phổ biến có thể được chia ra một cách đại khái thành hai loại: **Thuật toán dựa trên Hồi quy** và **Thuật toán dựa trên Phân đoạn**. Cũng có một số thuật toán kết hợp cả hai loại này. Các thuật toán dựa trên hồi quy lấy cảm hứng từ các thuật toán phát hiện đối tượng chung, thực hiện việc hồi quy hộp phát hiện bằng cách đặt các anchor, hoặc thậm chí trực tiếp thực hiện hồi quy điểm ảnh. Loại phương pháp này hoạt động tốt trong việc phân biệt văn bản có hình dạng đều đặn, nhưng kém trong việc phát hiện văn bản có hình dạng không đều. Ví dụ, CTPN tốt trong việc nhận dạng văn bản ngang, nhưng kém trong việc

Hình 3: Ví dụ về nhiệm vụ phát hiện văn bản

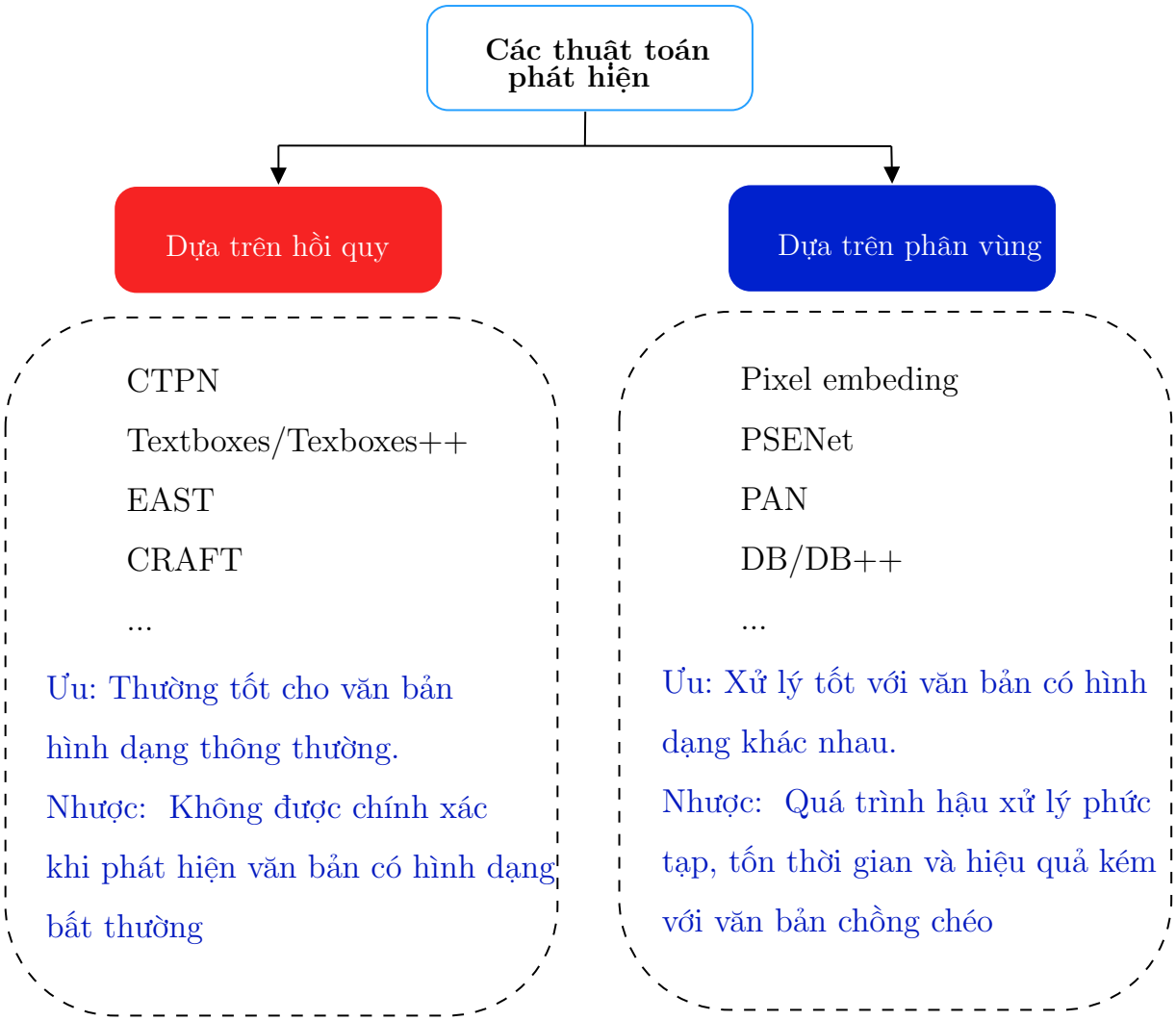
phát hiện văn bản uốn cong và xoắn. SegLink phù hợp hơn với văn bản dài, nhưng không thích hợp cho việc phát hiện văn bản phân tán thưa thớt. Các thuật toán dựa trên phân đoạn giới thiệu Mask-RCNN, loại thuật toán này có thể hoạt động tốt hơn trong việc phát hiện trong các tình huống và văn bản có các hình dạng khác nhau, nhưng hạn chế là việc xử lý sau cùng phức tạp, vì vậy có thể chậm về tốc độ và không thể phát hiện được văn bản chồng lấn [10].

2.3.2 Nhận dạng văn bản

Nhận dạng văn bản là việc nhận biết nội dung văn bản trong hình ảnh, và đầu vào thường là từ phần vùng chứa văn bản của ảnh được cắt ra bằng hộp văn bản được tạo ra từ việc phát hiện văn bản. Nhận dạng văn bản có thể được chia thành hai loại chính: **Nhận dạng Văn bản Đều đặn** và **Nhận dạng Văn bản Không đều đặn** dựa trên đường viền của văn bản cần nhận dạng.

Văn bản đều đặn chủ yếu đề cập đến các phông chữ in, văn bản được quét, và các nguồn tương tự có hướng chính đều. Văn bản không đều đặn thường không nằm trong tư thế ngang, thường uốn cong, bị che khuất và mờ mờ. Các tình huống văn bản không đều đặn thách thức rất lớn, và đó cũng là hướng nghiên cứu chính trong việc nhận dạng văn bản.

Các thuật toán nhận dạng văn bản đều đặn có thể được chia thành hai loại dựa trên các phương pháp giải mã khác nhau: Thuật toán dựa trên CTC và Thuật toán dựa trên Sequence2Sequence. Chúng khác nhau trong cách



Hình 4: Tổng quan về thuật toán phát hiện văn bản

chuyển đổi các đặc trưng chuỗi mà mạng học học được thành kết quả nhận dạng cuối cùng. Một ví dụ đại diện cho thuật toán dựa trên CTC là CRNN cổ điển.

Các thuật toán nhận dạng cho văn bản không đều đặn phong phú hơn. Các phương pháp như STAR-Net sửa chữa đường viền của văn bản không đều đặn thành các hình chữ nhật đều đặn bằng cách thêm các mô-đun sửa chữa như TPS trước khi thực hiện việc nhận dạng. Các phương pháp dựa trên Attention như RARE chú trọng hơn đến mối quan hệ giữa các phần trong chuỗi. Các phương pháp dựa trên phân đoạn xử lý mỗi ký tự trên dòng văn bản như một đơn vị riêng lẻ, làm cho việc nhận dạng ký tự đã phân đoạn dễ dàng hơn so với việc nhận dạng toàn bộ dòng văn bản sau khi sửa chữa. Ngoài ra, với sự phát triển nhanh chóng của Transformer và hiệu quả đã được xác minh trong các nhiệm vụ khác nhau trong những năm gần đây, nhiều thuật toán nhận dạng văn bản dựa trên transformer đã phát triển mạnh mẽ. Loại giải pháp này sử dụng cấu trúc transformer để giải quyết việc mô hình hóa sự phụ thuộc lâu dài trong CNN và đã đạt được kết quả tốt.

Hình 5: Văn bản đều đặn (trái) và Văn bản không đều đặn (phải)

2.3.3 Nhận dạng cấu trúc tài liệu

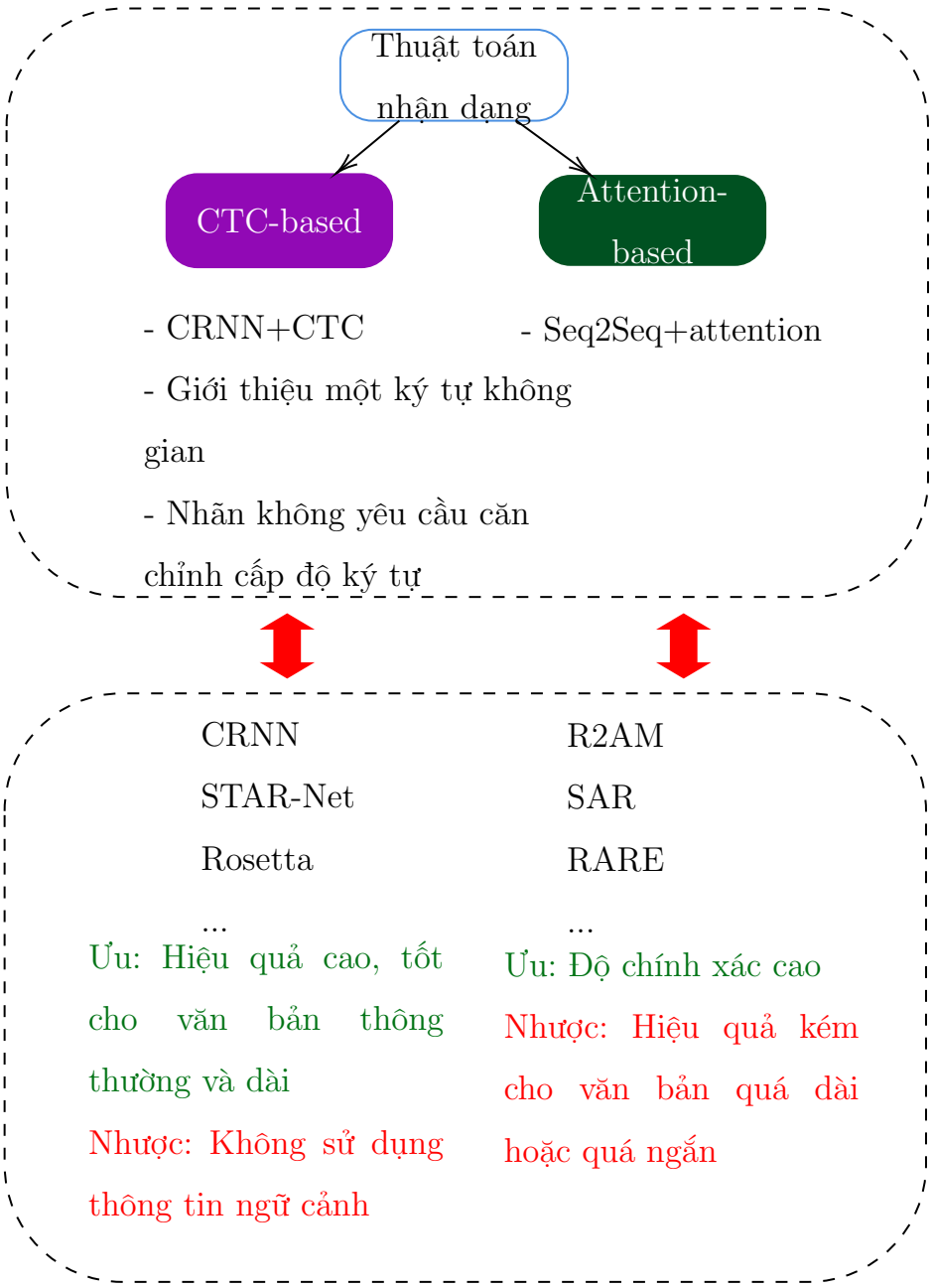
Công nghệ OCR có thể đáp ứng yêu cầu về phát hiện và nhận dạng văn bản. Tuy nhiên, trong các tình huống thực tế, điều chúng ta thường cần là thông tin có cấu trúc, chẳng hạn như trích xuất thông tin từ thẻ ID và hóa đơn, xác định có cấu trúc của bảng, và vân vân. Các tình huống ứng dụng của công nghệ OCR chủ yếu là trích xuất tài liệu nhanh, so sánh nội dung hợp đồng, so sánh thông tin tài chính trên các tài liệu cần thanh toán, và xác định tài liệu vận chuyển. Kết quả OCR + xử lý sau cùng là một kế hoạch cấu trúc thường được sử dụng, nhưng phức tạp và cần thiết phải được thiết kế cẩn thận, và thiếu sự tổng quát. Với sự phát triển liên tục của công nghệ OCR và nhu cầu về trích xuất thông tin có cấu trúc đang gia tăng, các công nghệ liên quan đến phân tích tài liệu thông minh, như phân tích bố cục, nhận dạng bảng, và trích xuất thông tin quan trọng, đã nhận được sự chú ý ngày càng tăng.

2.3.3.1 Phân tích bố cục

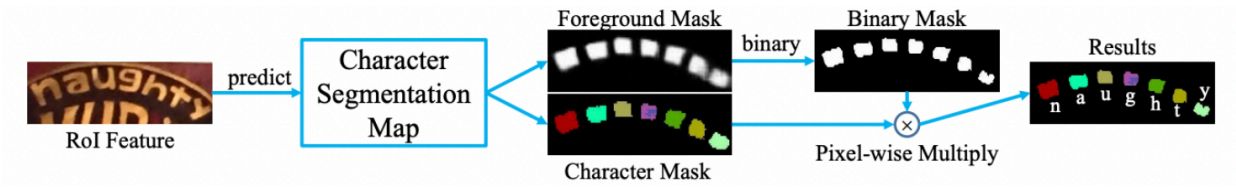
Phân tích bố cục được thực hiện để phân loại nội dung của hình ảnh tài liệu thành các loại như văn bản thuần túy, tiêu đề, bảng biểu, hình ảnh, v.v. Các phương pháp hiện tại thường thực hiện việc phát hiện hoặc phân đoạn chúng một cách riêng biệt. Ví dụ, Soto Carlos sử dụng thông tin ngữ cảnh và vị trí tự nhiên của nội dung tài liệu để cải thiện hiệu suất phát hiện vùng dựa trên thuật toán phát hiện mục tiêu Faster R-CNN. Sarkar Mausoom và đồng nghiệp đề xuất một cơ chế phân đoạn dựa trên tiên biết để huấn luyện mô hình phân đoạn tài liệu với các hình ảnh có độ phân giải cao, giải quyết vấn đề rằng các cấu trúc khác nhau trong các khu vực dày đặc không thể phân biệt và hợp nhất do việc giảm quá mức của hình ảnh gốc.

2.3.3.2 Nhận dạng bảng

Nhận dạng bảng là việc xác định và chuyển thông tin bảng của tài liệu thành một tệp Excel. Có nhiều loại và phong cách bảng khác nhau trong hình ảnh văn bản, chẳng hạn như các hàng và cột khác nhau và các loại văn bản khác nhau. Ngoài ra, phong cách của tài liệu và môi trường ánh sáng khi



Hình 6: CTC-based recognition algorithm VS. Attention-based recognition algorithm



Hình 7: Thuật toán nhận dạng dựa trên phân vùng ký tự

chụp ảnh đã đặt ra những thách thức lớn cho việc nhận dạng bảng, làm cho việc nhận dạng bảng trở thành một vấn đề nghiên cứu khó khăn trong việc hiểu tài liệu. Có nhiều phương pháp nhận dạng bảng. Ví dụ, vào những ngày đầu tiên, có các thuật toán truyền thống dựa trên các quy tắc heuristics, như thuật toán T-Rect được đề xuất bởi Kieninger và cộng sự, thường sử dụng quy tắc thiết kế thủ công và phát hiện và phân tích miền kết nối. Trong những năm gần đây, khi học sâu tiếp tục phát triển, một số thuật toán nhận dạng cấu trúc bảng dựa trên mạng CNN đã xuất hiện, như DeepTabStR được đề xuất bởi Siddiqui Shoaib Ahmed và cộng sự và TabStruct-Net được đề xuất bởi Raja Sachin và cộng sự. Ngoài ra, với sự gia tăng của Mạng Neural Đồ thị, một số nhà nghiên cứu đã thử áp dụng Mạng Neural Đồ thị vào việc nhận dạng cấu trúc bảng và coi việc nhận dạng bảng như một vấn đề tái tạo đồ thị dựa trên Mạng Neural Đồ thị. Đây là cách mà TGRNet được đề xuất bởi Xue Wenyuan và cộng sự hoạt động. Hơn nữa, có các giải pháp end-to-end có kết quả đầu ra cấu trúc bảng dưới dạng HTML bằng mạng. Hầu hết trong số này áp dụng Seq2Seq để dự đoán cấu trúc bảng như những thuật toán dựa trên attention hoặc transformer, bao gồm TableMaster.

2.3.3.3 Trích xuất thông tin chính

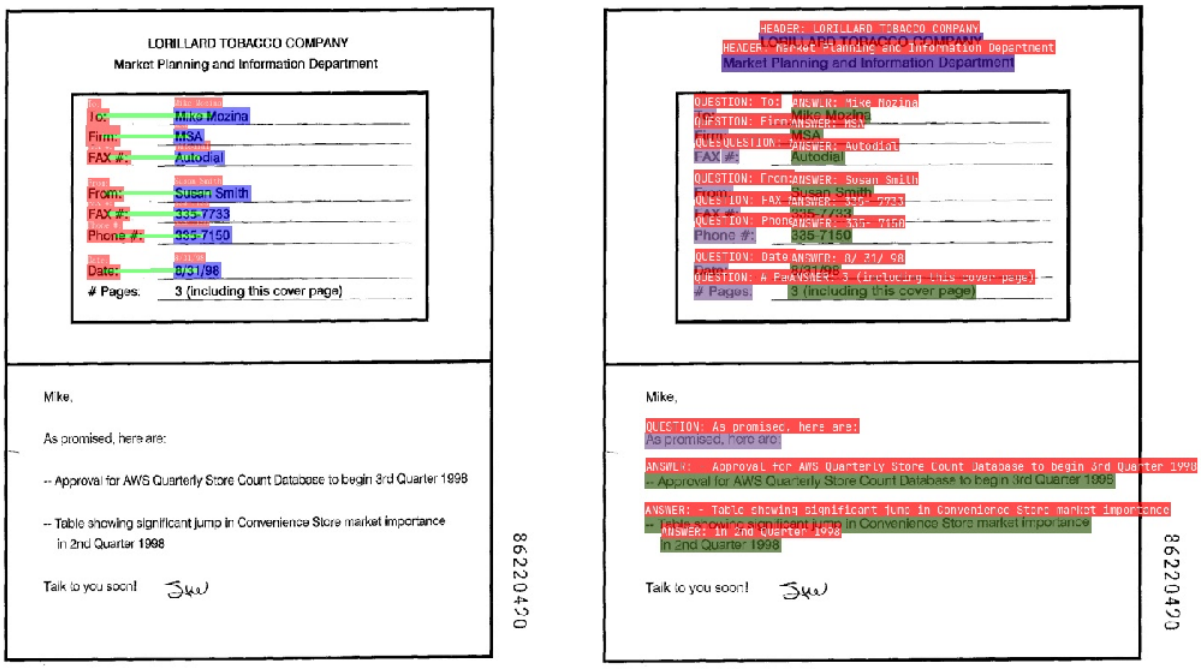
Trích xuất thông tin quan trọng (KIE) là một nhiệm vụ quan trọng trong Hỏi và Trả lời Văn bản (Document VQA). Nó liên quan đến việc trích xuất thông tin cần thiết từ hình ảnh, chẳng hạn như tên và số ID từ thẻ ID. Thông tin như vậy thường được xác định trong một nhiệm vụ, nhưng khác nhau giữa các nhiệm vụ khác nhau.

KIE thường được chia thành hai phần nhiệm vụ con để nghiên cứu (Hình 8):

- SER: Đây là việc nhận dạng thực thể ngữ nghĩa, phân loại từng đoạn văn bản được phát hiện. Ví dụ, nó chia văn bản thành tên và số thẻ ID như hình dưới đây.
- RE: Đây là việc trích xuất mối quan hệ, phân loại từng đoạn văn bản. Ví dụ, nó có thể phân loại văn bản thành câu hỏi và câu trả lời, sau đó tìm câu trả lời tương ứng cho mỗi câu hỏi. Như hình dưới đây, các hộp đỏ và đen đại diện cho câu hỏi và câu trả lời tương ứng, và các mũi tên màu vàng chỉ sự tương ứng giữa câu hỏi và câu trả lời.

Phương pháp KIE thông thường được phát triển dựa trên nhận dạng thực thể đặt tên (NER), nhưng loại phương pháp này chỉ sử dụng thông tin văn bản trong hình ảnh mà không sử dụng thông tin hình ảnh và cấu trúc. Do đó, nó không đạt độ chính xác cao. Trong những năm gần đây, nhiều giải pháp đã bắt đầu kết hợp thông tin hình ảnh và cấu trúc với thông tin văn bản. Do sử dụng các nguyên tắc khác nhau trong việc kết hợp thông tin đa tầng, các phương pháp này có thể được chia thành bốn loại:

- Phương pháp dựa trên lưới



Hình 8: RE(trái) VS SER(phải)

- Phương pháp dựa trên token
- Phương pháp dựa trên Graph Convolutional Network
- Phương pháp end-to-end

CHƯƠNG 3: PHƯƠNG PHÁP THỰC HIỆN

REFERENCES

[1] FPT.AI. "Nâng cấp quy trình số hóa tài liệu của doanh nghiệp với công nghệ ocr." (2020), [Online]. Available: https://fpt.ai/vi/nang-cap-quy-trinh-so-hoa-tai-lieu-cua-doanh-nghiep-voi-cong-nghe-ocr (visited on 07/08/2023).

[2] AWS. "Ocr (nhận dạng ký tự quang học) là gì?" (), [Online]. Available: https://aws.amazon.com/vi/what-is/ocr/ (visited on 07/08/2023).

[3] A. Singh, K. Bacchuwar, and A. Bhasin, "A survey of ocr applications," *International Journal of Machine Learning and Computing (IJMLC)*, Jan. 2012. DOI: 10.7763/IJMLC.2012.V2.137.

[4] Wikipedia. "Optical character recognition." (2002), [Online]. Available: https://en.wikipedia.org/wiki/Optical_character_recognition (visited on 07/08/2023).

[5] D. V. Everen. "The history of ocr." (2023), [Online]. Available: https://www.veryfi.com/ocr-api-platform/history-of-ocr/ (visited on 07/08/2023).

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[7] AWS. "Học sâu là gì?" (), [Online]. Available: https://aws.amazon.com/vi/what-is/deep-learning/ (visited on 10/08/2023).

[8] Wikipedia. "Deep learning." (2012), [Online]. Available: https://en.wikipedia.org/wiki/Deep_learning#Neural_networks (visited on 10/08/2023).

[9] J. W. Lum. "Ai vs traditional ocr." (2020), [Online]. Available: https://medium.com/staple-ai/artificial-intelligence-vs-traditional-ocr-how-staple-rides-the-wave-of-emerging-technologies-e60f2d295a26 (visited on 12/08/2023).

[10] C. Li, W. Liu, R. Guo, *et al.*, *Dive into OCR*. 2022, https://paddleocr.bj.bcebos.com/ebook/Dive_into_OCR.pdf.