BAN CƠ YẾU CHÍNH PHỦ

HỌC VIỆN KỸ THUẬT MẬT MÃ

ĐỀ CƯƠNG ĐỒ ÁN TỐT NGHIỆP

XÂY DỰNG HỆ THỐNG TRÍCH XUẤT VÀ TỔNG HỢP THÔNG TIN TỪ CÁC FILE PDF TRONG HỒ DỮ LIỆU

Ngành: Công nghệ thông tin

Mã số:

Sinh viên thực hiện:

Nguyễn Văn Đức Anh

Mã Sinh Viên: CT040304

Lớp: CT4C

Người hướng dẫn:

TS. Phạm Văn Hưởng

Khoa công nghệ thông tin – Học viện Kỹ thuật mật mã

Hà Nội, 2023

I. MỞ ĐẦU

1. Tính cấp thiết của đề tài

Ngày nay, công nghệ thông tin ngày càng phát triển và internet đã trở nên phổ biến, nhu cầu lưu trữ và chia sẻ dữ liệu của người dùng ngày càng lớn. File PDF cho phép hiển thị thông tin trên các thiết bị hay hệ điều hành khác nhau mà vẫn giữ nguyên định dạng và bố cục, tính bảo mật cao làm cho nó trở thành định dạng phổ biến cho việc lưu trữ và chia sẻ tài liệu trong nhiều lĩnh vực khác nhau như tài chính, giáo dục, y tế, v.v. Do đó, thông tin đươc lưu trữ trong các file PDF là rất lớn, đa dạng và cực kì hữu ích.

Bài toán trích xuất dữ liệu từ các file PDF làm tối ưu hóa các quy trình làm việc, tự động hóa việc trích xuất dữ liệu và giảm thiểu công sức của con người.

Mạng nơ-ron tích chập (Convolutional Neural Network - CNN) có thể dùng để học các đặc trưng của các cấu trúc file PDF như các trang, các đoạn văn, các bảng và các ô trong bảng, v.v. Các mô hình học sâu cho thấy hiệu quả cao trong các bài toán nhận dạng hình ảnh. Nhận thấy đây là phương án khả thi nên đồ án tập trung vào xây dựng mô hình CNN cho bài toán trích xuất dữ liệu từ file PDF.

2. Mục tiêu nghiên cứu của đề tài

+ Xây dựng kho dữ liệu PDF

+ Xây dựng hệ thống có khả năng trích xuất dữ liệu từ file PDF

+ Tổng hợp và trực quan hóa dữ liệu

3. Đối tượng và phạm vi nghiên cứu

+ Đối tượng: Mạng Nơ-ron tích chập, đặc trưng của tài liệu PDF.

+ Phạm vi: Nghiên cứu, xây dựng mô hình CNN giúp nhận dạng văn bản, bảng biểu, trích xuất thông tin, tổng hợp và trực quan hóa dữ liệu từ file PDF trong hồ dữ liệu.

4. Các nhiệm vụ chính cần thực hiện

Nội dung nghiên cứu được tập trung vào các nội dung chính như sau:

- Xây dựng kho lưu trữ, thu thập dữ liệu.
- Cơ sở lý thuyết về mạng nơ-ron tích chập.
- Xây dựng mô hình nhận dạng và trích xuất dữ liệu từ file PDF.
- Thử nghiệm, đánh giá và cải tiến.
- Tổng hợp và trực quan hóa dữ liệu

- Cài đăt, thử nghiệm, đánh giá hệ thống

5. Kết quả dự kiến

+ Lý thuyết:

Xây dựng mô hình CNN trích xuất, tổng hợp và trực quan hóa dữ liệu từ file PDF.

+ Thực nghiệm:

Ứng dụng di động cho phép người dùng chụp ảnh và chuyển hình ảnh thành file PDF, nhận dạng và trích xuất thông tin, trực quan háo dữ liệu.

II. DỰ KIẾN CÁC CHƯƠNG MỤC

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

DANH MỤC CÁC BẢNG BIỂU

DANH MỤC CÁC HÌNH VẼ

LỜI CẢM ƠN

MỞ ĐẦU

CHƯƠNG 1. TỔNG QUAN VỀ BÀI TOÁN TRÍCH XUẤT THÔNG TIN TỪ FILE PDF

1.1. Tổng quan về bài toán nhận dạng văn bản, bảng biểu

1.4. Các phương pháp nhận dạng và trích xuất thông tin từ file PDF

1.7. Tổng kết chương

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT VỀ MẠNG NƠ-RON TÍCH CHẬP

2.1. Mạng nơ-ron tích chập

2.2. Phân tích cấu trúc và bộ tham số của mạng nơ-ron tích chập

2.3. Mô hình ứng dụng mạng nơ-ron tích chập trong trích xuất thông tin từ file PDF

2.5. Tổng kết chương

CHƯƠNG 3. XÂY DỰNG HỆ THỐNG TRÍCH XUẤT VÀ TỔNG HỢP THÔNG TIN TỪ CÁC FILE PDF TRONG HỒ DỮ LIỆU.

3.1. Mô hình mạng nơ-ron tích chập cho bài toán OCR

3.2. Mô hình mạnh nơ-ron tích chập cho bài toán nhận dạng bảng biểu

3.3. Xây dựng mô hình

3.3. Phân tích, đánh giá mô hinh

3.4. Thực nghiệm và đánh giá kết quả

3.5. Tổng kết chương

KẾT LUẬN

TÀI LIỆU THAM KHẢO

PHỤ LỤC

III. TÀI LIỆU THAM KHẢO ĐỂ XÂY DỰNG ĐỀ CƯƠNG

[1] Phan Huy Hoang, "*[Deep Learning] Table Recognition - Simple is better than complex - Bài toán tái cấu trúc dữ liệu bảng biểu với deep learning*", Viblo.asia, 2022, Online: https://viblo.asia/p/deep-learning-table-recognition-simple-is-better-than-complex-bai-toan-tai-cau-truc-du-lieu-bang-bieu-voi-deep-learning-Qbq5QBYLKD8

[2] Tiep Vu, "*Machine Learning cho dữ liệu dạng bảng*", machinelearningcoban.com, Online: https://machinelearningcoban.com/tabml_book/intro.html

[3] "*Thuật toán CNN là gì? Cấu trúc mạng Convolutional Neural Network*", Topdev.vn, Online: https://topdev.vn/blog/thuat-toan-cnn-convolutional-neural-network/#cnn-convolutional-neural-network-la-gi

[4] Nguyen Tien Su, *"Giới thiệu bài toán OCR"*, tiensu.github.io, Online: https://tiensu.github.io/blog/63_ocr_introduction/

IV. KẾ HOẠCH THỰC HIỆN

STT	Thời gian	Nội dung thực hiện	Kết quả dự kiến

Hà Nội, ngày …. tháng …. năm …..

CÁN BỘ HƯỚNG DẪN	SINH VIÊN
TS. Trần Văn A	**Họ tên sinh viên**