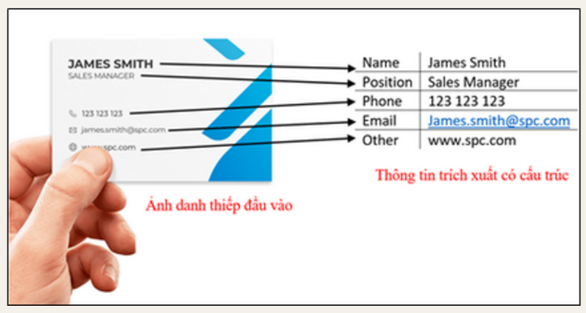
ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI
HANOI UNIVERSITY OF INDUSTRY

Sinh viên: Đức Minh Hoàng
GVHD: TS. Nguyễn Mạnh Cường

XÂY DỰNG HỆ THỐNG TRÍCH XUẤT THÔNG TIN DANH THIẾP BẰNG PADDLEOCR KẾT HỢP VỚI NER VÀ TINH CHỈNH KẾT QUẢ BẰNG LLM

Giới thiệu

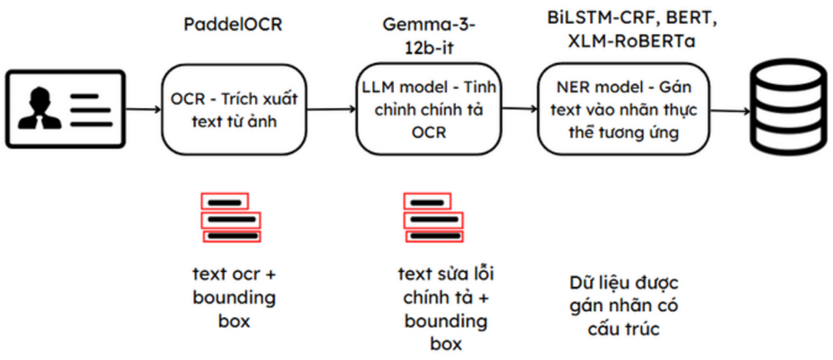
Trong thời đại số hóa, việc tự động trích xuất thông tin từ hình ảnh như danh thiếp đóng vai trò quan trọng trong quản lý dữ liệu cá nhân và doanh nghiệp. Tuy nhiên, danh thiếp thường có bố cục không chuẩn, sử dụng đa ngôn ngữ và chất lượng hình ảnh không đồng đều, gây khó khăn cho việc nhận dạng và phân loại thông tin. Các hệ thống OCR truyền thống tuy có thể trích xuất văn bản nhưng chưa đảm bảo độ chính xác cao, đặc biệt với tiếng Việt. Do đó, việc xây dựng một hệ thống kết hợp giữa OCR, sửa lỗi bằng mô hình ngôn ngữ lớn (LLM), và nhận dạng thực thể có tên (NER) sẽ góp phần nâng cao hiệu quả và độ tin cậy trong quá trình trích xuất thông tin.



Bài toán

- Tiếp nhận hình ảnh danh thiếp.
- Trích xuất chính xác các thông tin có cấu trúc như: Họ tên, Email, Số điện thoại, Địa chỉ, Chức vụ,...
- Tăng độ chính xác bằng cách khử nhiễu văn bản đầu ra từ OCR thông qua mô hình ngôn ngữ lớn.
- Triển khai hệ thống dưới dạng ứng dụng web có giao diện thân thiện, cho phép người dùng chỉnh sửa và quản lý danh thiếp.

Phương pháp tiếp cận



Hệ thống được thiết kế theo kiến trúc Pipeline gồm ba giai đoạn chính:

- Nhận diện văn bản (OCR)
  - Sử dụng PaddleOCR để trích xuất văn bản từ hình ảnh danh thiếp. PaddleOCR hỗ trợ tốt tiếng Việt và có độ chính xác cao trên nhiều định dạng.
- Sửa lỗi chính tả (Spell Correction)
  - Áp dụng Gemma-3-12B-IT, một mô hình ngôn ngữ lớn (LLM) nhằm sửa lỗi chính tả và cấu trúc câu do ảnh hưởng của quá trình OCR.
- Nhận dạng thực thể có tên (NER)
  - Triển khai và fine-tune XLM-RoBERTa, mô hình tiền huấn luyện đa ngôn ngữ, để gán nhãn cho các thành phần thông tin quan trọng.
  - So sánh hiệu quả với hai mô hình khác:
    - BiLSTM-CRF: truyền thống, phụ thuộc vào handcrafted features.
    - BERT: mạnh với tiếng Anh, nhưng hiệu suất chưa tối ưu với văn bản ngắn, đa ngôn ngữ.

Thực nghiệm và kết quả

Dữ liệu thực nghiệm:

- Synthetic data format theo dạng Json theo chuẩn IOB (Inside-Outside-Beginning)
- Với 6000 cặp dữ liệu huấn luyện trên 100 trường hợp khác nhau
- Gồm 16 nhãn I-B của: Name, Position, Company, Address, Phone, Email, Department
- Và một nhãn O

Finetuning các pretrained models trên NER có sẵn trên HuggingFace

- PassbyGrocer/bert_bilstm_crf -ner-weibo
- dslim/bert-base-NER
- Davlan/xlm-roberta-base-ner-hrl

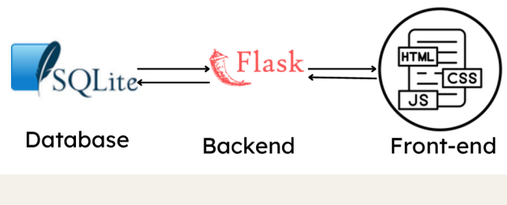
Các tham số huấn luyện

- Epoch: 100
- Batch size: 8
- Learning rate: 2e-5
- Weight decay: 0.01
- Tối ưu hóa: AdamW

Bảng so sánh kết quả trung bình từ k = 10 fold của ba mô hình BiLSTM-CRF, BERT và XLM-RoBERTa

Mô hình	Validation Loss	F1-scroce	Precision	Recall	Accuracy
BiLSTM-CRF	0.767s	0.593	0.553	0.553	0.872
BERT	0.771	0.612	0.557	0.683	0.869
XLM-RoBERTa	**0.555**	**0.844**	**0.821**	**0.867**	**0.949**

Luồng ứng dụng



Kết luận

- Đã xây dựng thành công hệ thống trích xuất thông tin danh thiếp tự động
- Kết hợp hiệu quả giữa thị giác máy tính (OCR), nhận xác thực thể có tên (NER) và LLM
- Ứng dụng có tiềm năng triển khai thực tế

Hạn chế:

- LLM còn nặng chưa tối ưu hóa chạy trên local (CPU)
- Chưa có hiệu chỉnh xoay với ảnh nghiêng

Định hướng:

- Xây dựng mô hình nhỏ gọn chính xác hóa chính tả từ OCR
- Bổ sung mô hình hiệu chỉnh xoay