ĐẠI HỌC QUỐC GIA TPHCM

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

23TNT1

Lab 2

Đề tài: Fine-tuning DeepSeek-OCR

Môn học: Nhập môn xử lý ngôn ngữ tự nhiên

Sinh viên thực hiện:

Lưu Thượng Hồng (23122006)

Giáo viên hướng dẫn:

PGS.TS. Đinh Điền

TS. Nguyễn Hồng Bửu Long

Ngày 21 tháng 12 năm 2025

Mục lục

1 Bối cảnh

1.1 Tổng quan về OCR và Thách thức với Tiếng Việt

Nhận dạng ký tự quang học (OCR - Optical Character Recognition) là công nghệ chuyển đổi hình ảnh chứa văn bản (tài liệu in, viết tay, ảnh chụp) thành định dạng văn bản máy tính có thể chỉnh sửa và tìm kiếm được.

Tuy nhiên, việc áp dụng OCR cho tiếng Việt gặp nhiều thách thức do đặc thù ngôn ngữ và chữ viết. Một số thách thức chính bao gồm:

- Đặc điểm ngôn ngữ: Tiếng Việt có nhiều dấu thanh và ký tự đặc biệt, điều này làm cho việc nhận diện trở nên khó khăn hơn so với các ngôn ngữ khác.
- Chất lượng hình ảnh: Các tài liệu quét có thể bị mờ, nghiêng hoặc có độ phân giải thấp, ảnh hưởng đến khả năng nhận diện của hệ thống OCR.
- Tính đa dạng của font chữ: Tiếng Việt sử dụng nhiều kiểu chữ khác nhau, từ chữ in đến chữ viết tay, điều này đòi hỏi hệ thống OCR phải được huấn luyện với một tập dữ liệu phong phú và đa dạng.

1.2 Tổng quan về mô hình DeepSeek-OCR

(Wei et al., 2025) DeepSeek-OCR là một mô hình Ngôn ngữ - Thị giác (Vision-Language Model - VLM) tiên tiến, được thiết kế theo hướng tiếp cận "nén ngữ cảnh quang học" (optical context compression). Thay vì chỉ nhận dạng ký tự đơn lẻ, mô hình xử lý toàn bộ hình ảnh tài liệu như một chuỗi token thị giác nén để giải mã ra văn bản.

1.2.1 Kiến trúc

Mô hình sử dụng kiến trúc End-to-End gồm hai thành phần chính nối tiếp nhau:

Encoder (DeepEncoder)

Đây là bộ mã hóa thị giác tùy chỉnh với khoảng 380M tham số, được thiết kế để xử lý ảnh độ phân giải cao nhưng vẫn tối ưu hóa bộ nhớ. DeepEncoder bao gồm 3 module con:

- Visual Perception: Sử dụng SAM-base (80M tham số) với cơ chế *window attention* để trích xuất các đặc trưng chi tiết cục bộ.
- Compressor: Một module tích chập (Conv layer) thực hiện giảm mẫu (downsample) 16 lần, giúp nén đáng kể số lượng vision tokens (ví dụ giảm từ 4096 xuống còn 256 tokens).
- Visual Knowledge: Sử dụng CLIP-large (300M tham số) với cơ chế *global attention* để nắm bắt ngữ nghĩa toàn cục và tri thức từ các token đã nén.

Decoder

Về decoder, DeepSeek-OCR sử dụng mô hình ngôn ngữ lớn DeepSeek3B-MoE (Mixture-of-Experts). Mặc dù có tổng cộng 3B tham số, mô hình chỉ kích hoạt khoảng 570M tham số trong quá trình suy luận, đảm bảo tốc độ xử lý nhanh và hiệu quả cao

1.2.2 Đầu vào

DeepSeek-OCR có khả năng xử lý linh hoạt các loại đầu vào thông qua cơ chế Đa phân giải (Multi-resolution support):

- Native Resolution: Hỗ trợ các chế độ phân giải cố định như Tiny, Small, Base và Large (tối đa 1280x1280 pixel).
- Dynamic Resolution (Gundam Mode): Hỗ trợ xử lý ảnh kích thước cực lớn hoặc tỷ lệ khung hình đặc biệt (như trang báo, tài liệu dài) bằng cách chia nhỏ ảnh (tiling) kết hợp với cái nhìn toàn cảnh, giúp không bị mất chi tiết.

1.2.3 Đầu ra

Mô hình hỗ trợ đa dạng định dạng đầu ra phục vụ nhiều mục đích khác nhau:

- Văn bản thuần & Markdown: Cho các tài liệu văn bản thông thường.
- HTML: Để nhận dạng và tái tạo cấu trúc bảng biểu (Tables) phức tạp.
- LaTeX / SMILES: Dành cho việc nhận dạng công thức toán học và công thức hóa học.
- Tọa độ (Bounding boxes): Cung cấp vị trí của đối tượng/văn bản cho các tác vụ grounding hoặc detection.

2 Phương pháp

Tài liệu

Wei, H., Sun, Y., & Li, Y. (2025). Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.