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 28 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.

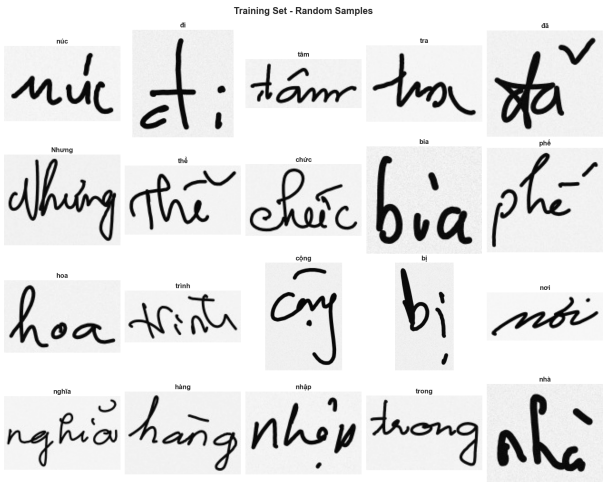
2 Phương pháp

2.1 Bộ dữ liệu: UIT-HWDB-word

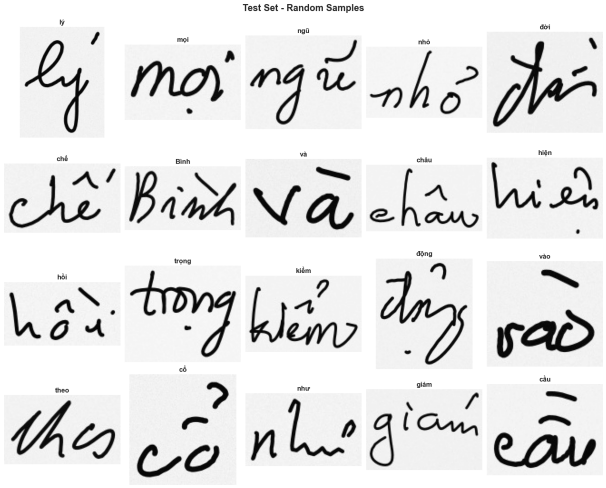
2.1.1 Tổng quan bộ dữ liệu

Ta sử dụng bộ dữ liệu UIT-HWDB-word, một phần của bộ dữ liệu chữ viết tay tiếng Việt UIT-HWDB (Nguyen et al., 2022).

Một số mẫu chữ viết tay từ bộ dữ liệu được minh họa trong Hình 3.



Hình 1: (a) Tập huấn luyện



Hình 2: (b) Tập kiểm thử

Hình 3: Một số mẫu chữ viết tay từ bộ dữ liệu UIT-HWDB-word

2.1.2 Cấu trúc dữ liệu

Bộ dữ liệu được tổ chức thành hai tập chính: tập huấn luyện (train) và tập kiểm tra (test). Cấu trúc thư mục như sau:

```
UIT_HWDB_word/
    train_data/
        1/
            image_001.png
            ...
            label.json
        ...
```

```
test_data/
    250/
        ...
        label.json
    ...
```

Mỗi thư mục con chứa các ảnh chữ viết tay và một file `label.json` chứa nhãn tương ứng cho từng ảnh. Định dạng của file nhãn như sau:

```
{
    "1.jpg": "xin chào",
    "2.jpg": "cảm ơn",
    ...
}
```

2.2 Phân tích dữ liệu (Exploratory Data Analysis)

Quá trình phân tích khám phá dữ liệu (EDA) được thực hiện chi tiết trên tập huấn luyện (107,607 mẫu) và tập kiểm tra (2,881 mẫu) nhằm định hướng cho các quyết định tiền xử lý và cấu hình mô hình.

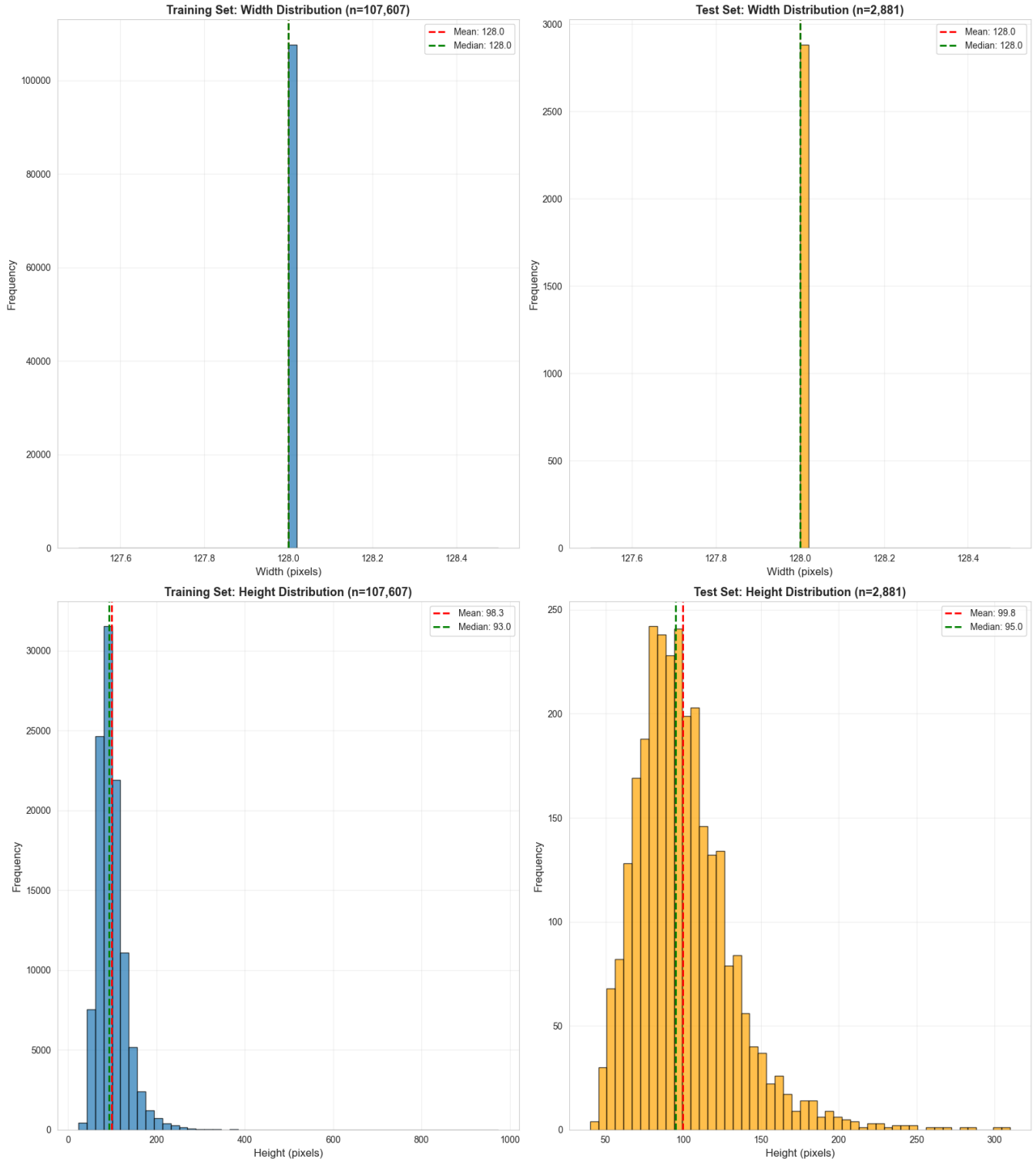
2.2.1 Đặc điểm hình ảnh

Phân tích thống kê kích thước ảnh cho thấy những đặc trưng quan trọng:

- Chiều rộng cố định: Tất cả các ảnh trong bộ dữ liệu đều có chiều rộng 128 pixels.
- Chiều cao biến thiên: Chiều cao ảnh có sự dao động lớn, từ 23 pixels đến 974 pixels, với giá trị trung bình khoảng 98 pixels và trung vị là 93 pixels.
- Điểm ngoại lai (Outliers): Mặc dù phần lớn ảnh có chiều cao dưới 128 pixels, sự tồn tại của các ảnh có chiều cao lên tới gần 1000 pixels cho thấy có những mẫu chữ viết tay rất dài hoặc được viết theo chiều dọc.

Từ kết quả này, việc lựa chọn kích thước ảnh đầu vào cho mô hình là 384x384 được đánh giá là tối ưu. Kích thước này đủ lớn để chứa trọn vẹn đa số các ảnh (với chiều cao trung bình ∼98px) mà

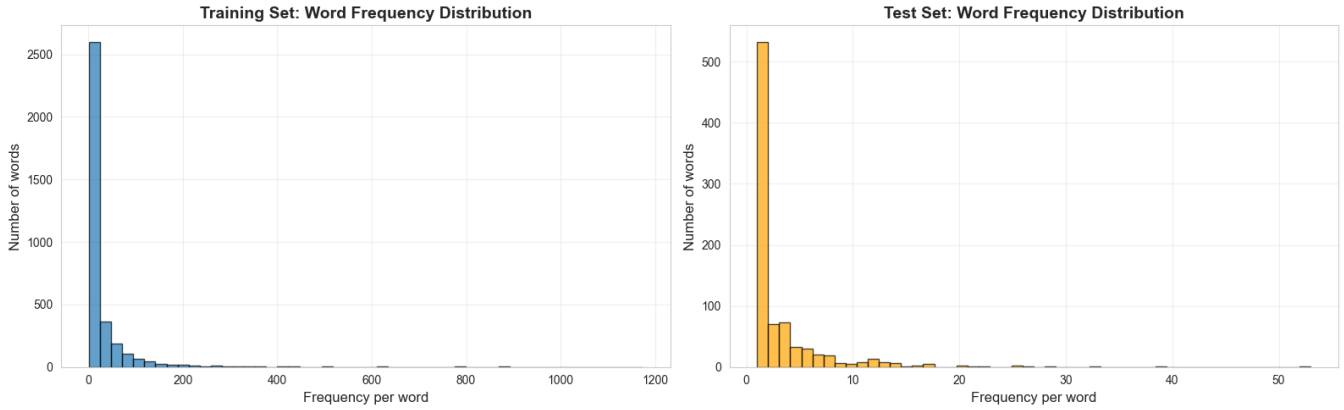
không cần co giãn quá nhiều, đồng thời hạn chế lãng phí tài nguyên tính toán cho phần padding của các ảnh nhỏ.



Hình 4: Phân bố kích thước ảnh trong bộ dữ liệu UIT-HWDB-word

2.2.2 Phân bố nhãn

- Phân phối Long-tail: Tần suất xuất hiện của các nhãn từ tuân theo quy luật phân phối đuôi dài. Một lượng nhỏ các từ thông dụng chiếm tỷ trọng lớn trong tập dữ liệu, trong khi có những từ chỉ xuất hiện rất ít lần.



Hình 5: Phân bố nhãn

2.3 Tiền xử lý dữ liệu (Data Preprocessing)

2.3.1 Chuẩn bị dữ liệu

Quá trình chuẩn bị dữ liệu bao gồm việc đọc các file `label.json` từ các thư mục con, tải ảnh sử dụng thư viện PIL và chuyển đổi sang định dạng RGB.

2.3.2 Định dạng hội thoại (Conversation Format)

Để fine-tune mô hình DeepSeek-OCR (một mô hình Vision-Language Model), dữ liệu được chuyển đổi sang định dạng hội thoại chuẩn:

```
{
    "messages": [
        {
            "role": "<|User|>",
            "content": "<image>\nFree OCR.",
            "images": [<PIL.Image>]
        },
```

```
        {
            "role": "<|Assistant|>",
            "content": "nhãn_của_ảnh"
        }
    ]
}
```

Trong đó, token `<image>` đại diện cho vị trí chèn các embedding của ảnh, và câu lệnh "Free OCR." đóng vai trò là prompt hướng dẫn mô hình thực hiện tác vụ nhận dạng ký tự quang học.

2.3.3 Xử lý ảnh (Image Processing)

Trong quá trình huấn luyện, ảnh đầu vào được xử lý thông qua `DeepSeekOCRDataCollator` với các tham số cấu hình như sau:

- **Kích thước ảnh (Image Size):** 384x384 pixels. Tham số này được lựa chọn dựa trên kết quả EDA cho thấy phần lớn ảnh trong bộ dữ liệu có kích thước nhỏ, do đó kích thước 384x384 là đủ để bao quát chi tiết ảnh mà không gây lãng phí tài nguyên tính toán.
- **Chế độ cắt (Crop Mode):** Tắt (`False`). Do kích thước ảnh gốc chủ yếu là nhỏ và phù hợp với kích thước đầu vào (384x384), việc sử dụng cơ chế dynamic cropping (thường dùng cho ảnh độ phân giải cao) là không cần thiết. Việc tắt chế độ này giúp đơn giản hóa quá trình xử lý và tối ưu hóa tốc độ huấn luyện.
- **Chuẩn hóa (Normalization):** Mean = (0.5, 0.5, 0.5), Std = (0.5, 0.5, 0.5).

Lưu ý rằng trong quá trình suy diễn (inference), kích thước ảnh và chế độ crop có thể được điều chỉnh (ví dụ: 1024x1024 và bật crop mode) để đạt độ chính xác cao hơn.

2.4 Kiến trúc mô hình và Fine-tuning

2.4.1 Mô hình cơ sở

Dự án sử dụng mô hình **DeepSeek-OCR**, được tải và tối ưu hóa thông qua thư viện **Unsloth** (`FastVisionModel`). Unsloth giúp tăng tốc độ huấn luyện và giảm bộ nhớ tiêu thụ thông qua các kỹ thuật tối ưu hóa kernel và lượng tử hóa.

2.4.2 Chiến lược Fine-tuning: LoRA

Để tinh chỉnh mô hình trên tập dữ liệu tiếng Việt với tài nguyên hạn chế, kỹ thuật **LoRA (Low-Rank Adaptation)** được áp dụng. Cấu hình LoRA cụ thể như sau:

- **Rank (r):** 16
- **Alpha:** 16
- **Dropout:** 0
- **Target Modules:** Áp dụng lên tất cả các lớp linear projection trong mô hình attention: `q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj`.

2.4.3 Tham số huấn luyện (Hyperparameters)

Quá trình huấn luyện được thực hiện với các tham số chính:

- **Số epochs:** 2
- **Batch size:** 32 (per device)
- **Gradient Accumulation Steps:** 2
- **Learning Rate:** 1e-4
- **Optimizer:** AdamW 8-bit (giúp tiết kiệm bộ nhớ VRAM)
- **Scheduler:** Linear decay với 5 bước warmup
- **Độ chính xác:** BF16 (nếu phần cứng hỗ trợ) hoặc FP16

2.5 Độ đo đánh giá (Evaluation Metrics)

Mô hình được đánh giá dựa trên các độ đo phổ biến trong bài toán OCR:

- **CER (Character Error Rate):** Tỷ lệ lỗi ký tự, đo lường khoảng cách chỉnh sửa (Levenshtein distance) giữa chuỗi dự đoán và nhãn thực tế, chuẩn hóa theo độ dài nhãn.
- **Accuracy:** Được tính toán dựa trên CER (1 - CER) hoặc tỷ lệ khớp chính xác (Exact Match) tùy ngữ cảnh phân tích.

- **Exact Match:** Tỷ lệ số mẫu dự đoán hoàn toàn chính xác so với nhãn gốc.

Tài liệu

Nguyen, N. H., Vo, D. T. D., & Nguyen, K. V. (2022). UIT-HWDB: using transferring method to construct A novel benchmark for evaluating unconstrained handwriting image recognition in vietnamese. *RIVF International Conference on Computing and Communication Technologies, RIVF 2022, Ho Chi Minh City, Vietnam, December 20-22, 2022*, 659–664. https://doi.org/10.1109/RIVF55975.2022.10013898

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