HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



NGUYỄN PHƯƠNG ANH

ỨNG DỤNG MÔ HÌNH THỊ GIÁC-NGÔN NGỮ LỚN VÀO BÀI TOÁN NHẬN DẠNG VÀ TRÍCH XUẤT THÔNG TIN TỪ ẢNH TÀI LIỆU TIẾNG VIỆT

Chuyên ngành: Khoa học máy tính
Mã số: 8.48.01.01

TÓM TẮT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

HÀ NỘI - NĂM 2025

Đề án tốt nghiệp được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: ……

(Ghi rõ học hàm, học vị)

Phản biện 1: ……

Phản biện 2: ……..

Đề án tốt nghiệp sẽ được bảo vệ trước Hội đồng chấm đề án tốt nghiệp thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng .... năm .....

Có thể tìm hiểu đề án tốt nghiệp tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

MỤC LỤC

MỤC LỤC.....2

CHƯƠNG 1: TỔNG QUAN VỀ BÀI TOÁN..... 3

1.1. Giới thiệu chung về bài toán OCR..... 3

1.2. Đặc thù tiếng Việt trong bài toán OCR..... 3

1.2.1. Hệ thống chữ viết và dấu tiếng Việt..... 3

1.2.2. Bố cục tài liệu và đa dạng nguồn dữ liệu..... 3

1.2.3. Chữ in và chữ viết tay..... 3

1.3. Các hướng tiếp cận OCR hiện nay.....4

1.3.1. Pipeline OCR truyền thống..... 4

1.3.2. OCR dựa trên học sâu.....5

1.4. Các nghiên cứu về OCR tiếng Việt.....5

1.4.1. Nhận dạng chữ in và văn bản in trong tài liệu.....5

1.4.2. Nhận dạng chữ trong ảnh cảnh.....5

1.4.3. Nhận dạng chữ viết tay tiếng Việt..... 5

1.5. Xu hướng mô hình thị giác–ngôn ngữ và OCR tài liệu tiếng Việt..... 5

1.6. Đánh giá chung và khoảng trống nghiên cứu..... 6

1.7. Kết luận chương.....6

CHƯƠNG 2: PHƯƠNG PHÁP ĐỀ XUẤT..... 7

2.1. Định hướng và mục tiêu của phương pháp.....7

2.2. Cơ sở lý thuyết.....7

2.2.1. Cơ sở lý thuyết về mô hình thị giác–ngôn ngữ (VLM)..... 7

2.2.2. Kiến trúc mô hình Qwen 2.5-VL 7B..... 8

2.2.3. Cơ sở lý thuyết về fine-tuning và LoRA.....9

2.3. Thiết kế dữ liệu huấn luyện cho Qwen 2.5-VL..... 9

2.3.1. Loại dữ liệu..... 9

2.3.2. Tiền xử lý và chuẩn hóa.....9
2.3.3. Định dạng dữ liệu theo phong cách hội thoại.....9
2.4. Pipeline tinh chỉnh Qwen 2.5-VL bằng LoRA.....10
2.4.1. Các bước tổng quát.....10
2.4.2. Cấu hình huấn luyện cụ thể.....11
2.5. Quy trình suy luận (inference) cho OCR.....11
2.6. Thiết lập benchmark và tiêu chí đánh giá.....12
2.7. Kết luận chương.....12
CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ.....13
3.1. Bộ dữ liệu thực nghiệm.....13
3.2. Độ đo đánh giá (Metrics).....13
3.3. Quy trình tinh chỉnh mô hình.....13
3.4. Kết quả thực nghiệm.....15
3.5. Xây dựng ứng dụng.....16
3.7. Kết luận chương.....17
KẾT LUẬN.....18

MỞ ĐẦU

Trong bối cảnh chuyển đổi số mạnh mẽ tại Việt Nam, nhu cầu số hóa tài liệu giấy và tự động hóa quy trình xử lý văn bản ngày càng trở nên cấp thiết đối với cả doanh nghiệp lẫn cơ quan nhà nước. Dù OCR đã được ứng dụng rộng rãi, các phương pháp truyền thống như Tesseract, PaddleOCR hoặc các mô hình CRNN/Transformer trong pipeline detection–recognition vẫn gặp nhiều hạn chế khi xử lý tiếng Việt do đặc thù hệ thống dấu thanh, sự đa dạng font chữ và chất lượng tài liệu thực tế không đồng nhất. Chữ viết tay tiếng Việt tiếp tục là một thách thức lớn do thiếu chuẩn hóa và biến thiên mạnh giữa các cá nhân.

Sự xuất hiện của các mô hình thị giác–ngôn ngữ lớn (VLM) như LayoutLM, Donut và đặc biệt là Qwen 2.5-VL mở ra hướng tiếp cận mới cho OCR, cho phép mô hình hiểu ngữ cảnh toàn cục của tài liệu và sinh văn bản trực tiếp mà không cần pipeline tách biệt. Qwen 2.5-VL có nhiều ưu thế trong xử lý ảnh tài liệu phức tạp, tuy nhiên vẫn cần được tinh chỉnh để phù hợp với đặc thù tiếng Việt.

Hiện nay, vẫn thiếu các nghiên cứu đánh giá có hệ thống về hiệu quả của VLM cho OCR tiếng Việt, đặc biệt trên dữ liệu thực tế và chữ viết tay. Khoảng trống này là động lực để thực hiện đề án "Ứng dụng mô hình thị giác–ngôn ngữ lớn vào bài toán nhận dạng và trích xuất thông tin từ ảnh tài liệu tiếng Việt".

Đề án tập trung xây dựng tập dữ liệu đa dạng gồm chữ in khó và chữ viết tay, chuẩn hóa theo định dạng hội thoại phù hợp với VLM, tinh chỉnh mô hình Qwen 2.5-VL bằng kỹ thuật LoRA và đánh giá hiệu năng trên bộ benchmark độc lập. Mô hình được so sánh với hai baseline mạnh: PaddleOCR và Vintern.

Cấu trúc đề án gồm ba chương:

- Chương 1: Tổng quan OCR tiếng Việt và các hướng tiếp cận hiện có, gồm cả VLM.

- Chương 2: Phương pháp đề xuất, pipeline tinh chỉnh Qwen 2.5-VL bằng LoRA và thiết kế dữ liệu.

- Chương 3: Thực nghiệm, đánh giá hiệu quả mô hình qua CER/WER và trình bày ứng dụng demo trích xuất thông tin từ ảnh tài liệu.

CHƯƠNG 1: TỔNG QUAN VỀ BÀI TOÁN

1.1. Giới thiệu chung về bài toán OCR

OCR là bài toán chuyển đổi văn bản trong ảnh thành dạng ký tự có thể xử lý tự động, đóng vai trò quan trọng trong số hóa tài liệu, chính phủ điện tử và tự động hóa doanh nghiệp.

Trong khi OCR tiếng Anh đã đạt độ chính xác rất cao trong môi trường chuẩn hóa, OCR tiếng Việt vẫn còn nhiều khó khăn do hệ thống dấu phức tạp, dữ liệu không chuẩn hóa và chất lượng ảnh đa dạng.

Các khảo sát DAR chỉ ra rằng OCR vẫn là "nút thắt" trong hệ thống xử lý tài liệu tiếng Việt, đặc biệt với dữ liệu thực tế như ảnh chụp, scan lệch hoặc văn bản viết tay.

1.2. Đặc thù tiếng Việt trong bài toán OCR

1.2.1. Hệ thống chữ viết và dấu tiếng Việt

Tiếng Việt có 29 chữ cái, kết hợp với nhiều dấu thanh và dấu phụ, tạo ra hơn ~150–170 biến thể ký tự khác nhau.

Dấu nhỏ và dễ mất khi ảnh mờ, thấp chất lượng, khiến mô hình khó phân biệt, đặc biệt giữa các nhóm hình dạng gần nhau (a/à/ả/ã/ạ; o/ô/ơ…).

1.2.2. Bố cục tài liệu và đa dạng nguồn dữ liệu

Nguồn tài liệu Việt Nam rất phong phú: giấy tờ hành chính, biểu mẫu, hóa đơn, sách, văn bản viết tay.

Chất lượng hình ảnh thay đổi mạnh (scan nghiêng, nhiễu, ảnh chụp điện thoại…), bố cục không chuẩn hoá → OCR phải thích ứng tốt để đạt độ chính xác cao.

1.2.3. Chữ in và chữ viết tay

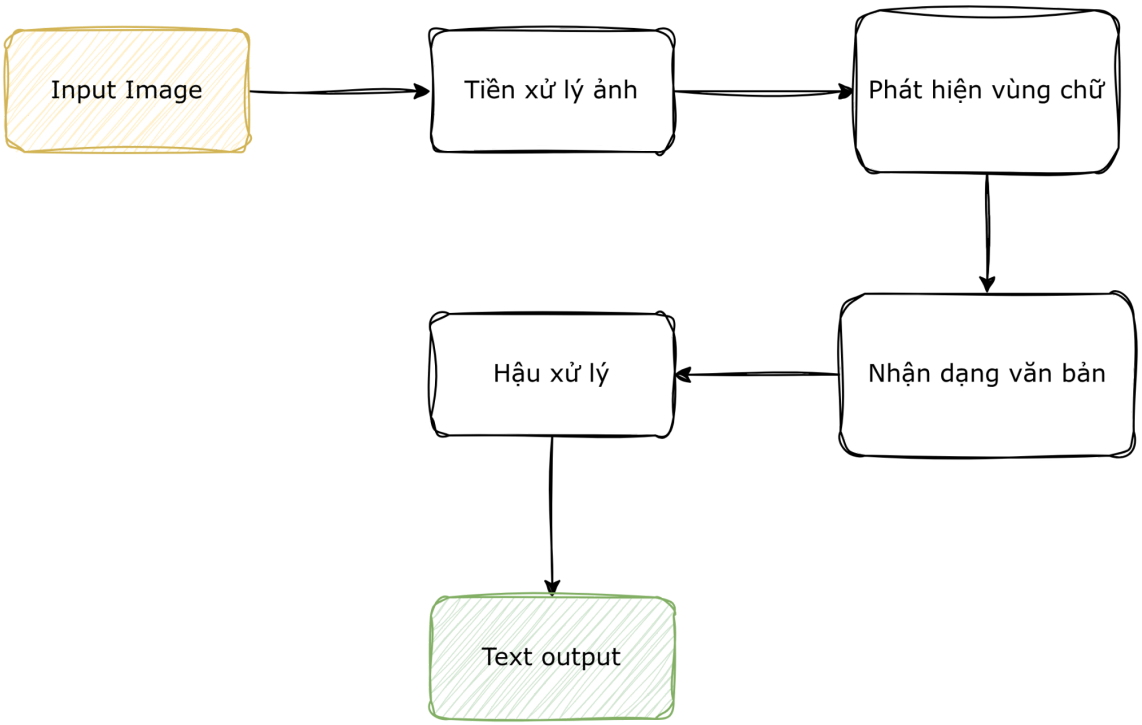
Chữ in tương đối ổn định và OCR có thể đạt kết quả tốt khi ảnh đủ sạch.

Chữ viết tay biến thiên lớn về nét chữ, cách ghi dấu, thói quen viết tắt và bố cục dòng, khiến nhận dạng khó hơn nhiều. Các cuộc thi (Kalapa) và dataset (VNOnDB, UIT-HWDB) cho thấy đây vẫn là bài toán thách thức.

1.3. Các hướng tiếp cận OCR hiện nay

1.3.1. Pipeline OCR truyền thống

1. Pipeline gồm: tiền xử lý → phát hiện vùng chữ → nhận dạng → hậu xử lý.
2. Tesseract dùng LSTM + CTC; PaddleOCR sử dụng detector DB/EAST và nhiều recognizer (CRNN, Transformer).
3. Ưu điểm: dễ triển khai, tùy biến tốt.
4. Nhược điểm: phụ thuộc mạnh vào text detection; dễ mất dấu tiếng Việt; không phù hợp tài liệu mờ, cong, bố cục phức tạp.



Hình 1.1: Pipeline OCR truyền thống

1.3.2. OCR dựa trên học sâu

Xu hướng chuyển sang mô hình end-to-end với CNN, RNN/LSTM, Transformer, CTC hoặc attention-based seq2seq.

Các mô hình tiêu biểu: CRNN, SATRN, Rosetta, TrBA.

Trong tiếng Việt, VietOCR (Transformer-based) là đại diện phổ biến, cho phép fine-tune theo domain.

1.4. Các nghiên cứu về OCR tiếng Việt

1.4.1. Nhận dạng chữ in và văn bản in trong tài liệu

Nhiều nghiên cứu sử dụng CNN/Transformer và các bộ dữ liệu như UIT-ViODC, ViTextVQA.
Cộng đồng dùng Tesseract, VietOCR, PaddleOCR theo pipeline tùy chỉnh cho tiếng Việt.

1.4.2. Nhận dạng chữ trong ảnh cảnh

- Dataset: VinText, VnSceneText.
- Áp dụng detector mạnh (EAST, DB…) kết hợp recognizer SATRN/CRNN.
- Thách thức lớn khi chữ nhỏ, mờ hoặc ánh sáng kém.

1.4.3. Nhận dạng chữ viết tay tiếng Việt

Dữ liệu viết tay còn hạn chế, độ biến thiên ký tự cao.
Các phương pháp sử dụng CRNN/CTC, CNN-LSTM hoặc Transformer.
Các cuộc thi và dataset (VNOnDB, Kalapa, Vietnamese Handwriting OCR) thúc đẩy nghiên cứu nhưng bài toán vẫn rất khó.

1.5. Xu hướng mô hình thị giác–ngôn ngữ và OCR tài liệu tiếng Việt

Mô hình VLM như LayoutLM, Donut, Pix2Struct, Qwen-VL cho phép đọc hiểu tài liệu không cần detection, nhờ sinh văn bản trực tiếp theo kiểu generative.

VLM hiểu bố cục toàn trang, hỗ trợ Document-VQA và trích xuất thông tin.

Qwen 2.5-VL nổi bật nhờ xử lý ảnh độ phân giải cao, hỗ trợ tiếng Việt tốt và dễ tinh chỉnh (LoRA/SFT).

Đề tài hướng đến tinh chỉnh Qwen 2.5-VL 7B và so sánh với các baseline như PaddleOCR và Vintern.

1.6. Đánh giá chung và khoảng trống nghiên cứu

Mặc dù OCR tiếng Việt đã có nhiều tiến bộ, hiệu suất trên dữ liệu thực tế vẫn chưa ổn định. VLM là hướng mới nhưng chưa có nhiều nghiên cứu thực nghiệm chi tiết cho tiếng Việt.
Khoảng trống: cần dataset "khó", pipeline tinh chỉnh VLM và đánh giá có hệ thống với baseline truyền thống.

1.7. Kết luận chương

Chương 1 trình bày đặc thù ngôn ngữ, đặc điểm dữ liệu, các phương pháp OCR và xu hướng VLM.

OCR tiếng Việt vẫn có nhiều hạn chế do dữ liệu phức tạp và hệ thống dấu đặc biệt.

VLM mở ra hướng tiếp cận mạnh mẽ, có khả năng vượt qua hạn chế của pipeline truyền thống.

Đây là cơ sở cho việc tinh chỉnh Qwen 2.5-VL 7B trong các chương tiếp theo.

CHƯƠNG 2: PHƯƠNG PHÁP ĐỀ XUẤT

2.1. Định hướng và mục tiêu của phương pháp

- Phương pháp đề xuất tận dụng mô hình thị giác–ngôn ngữ Qwen 2.5-VL 7B để thực hiện OCR theo hướng sinh văn bản trực tiếp từ ảnh, thay thế pipeline truyền thống nhiều bước.
- Mô hình được tinh chỉnh bằng LoRA trên tập dữ liệu tiếng Việt gồm chữ in khó và chữ viết tay nhằm nâng cao khả năng nhận dạng trong môi trường thực tế.
- Mục tiêu chính: cải thiện CER/WER, đánh giá ưu thế so với PaddleOCR, Vintern và Qwen 2.5-VL gốc.

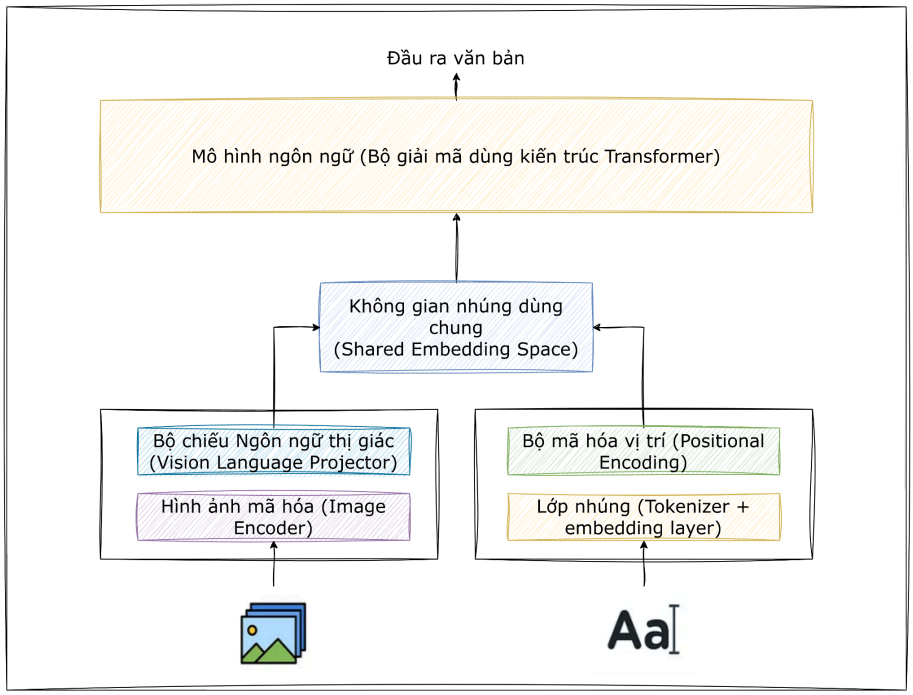
2.2. Cơ sở lý thuyết

2.2.1. Cơ sở lý thuyết về mô hình thị giác–ngôn ngữ (VLM)

VLM kết hợp visual encoder (CNN/ViT) và language model (BERT, GPT, T5…) thông qua cơ chế hợp nhất đa mô thức (projection, cross-attention).

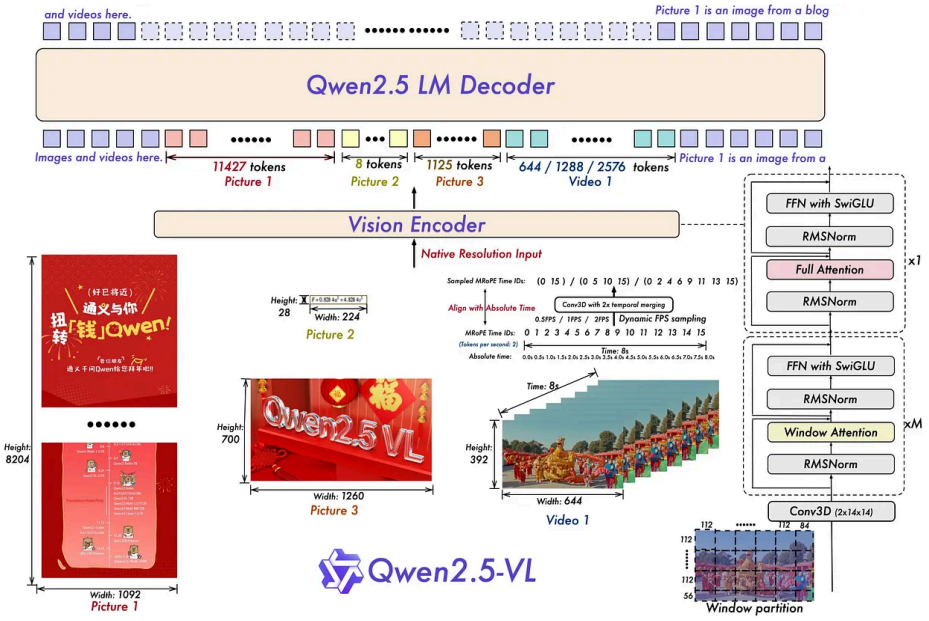
Mô hình tạo không gian biểu diễn chung cho ảnh và văn bản, giúp xử lý các tác vụ như OCR, captioning, VQA và document understanding.

Trong OCR, VLM cho phép đọc toàn trang và sinh nguyên văn nội dung mà không cần detection/cropping.



Hình 2.1: Kiến trúc chung của mô hình VLM

2.2.2. Kiến trúc mô hình Qwen 2.5-VL 7B



Hình 2.2: Kiến trúc mô hình Qwen2.5-VL (Nguồn: Tham khảo tại link: [Qwen2.5 VL](#))

Qwen 2.5-VL gồm bốn phần:

- Backbone ngôn ngữ Qwen 2.5 (decoder-only Transformer).
- Vision Transformer độ phân giải cao (high-resolution ViT).
- Projection layer căn chỉnh visual tokens → language token space.
- Multimodal cross-attention giúp mô hình hội tụ sâu hai mô thức.
  Mô hình hỗ trợ ảnh độ phân giải cao, đọc văn bản nhỏ, hiểu bố cục tài liệu, và sinh chuỗi OCR theo dạng generative.

2.2.3. Cơ sở lý thuyết về fine-tuning và LoRA

- Fine-tuning giúp mô hình thích ứng với domain mới, nhưng cập nhật toàn bộ tham số của VLM lớn rất tốn tài nguyên.
- LoRA là phương pháp PEFT phổ biến, chỉ học thêm các ma trận hạng thấp gắn vào lớp attention.
- LoRA giảm chi phí GPU, tăng tốc huấn luyện, giúp fine-tuning mô hình 7B trên GPU 24GB khả thi.

2.3. Thiết kế dữ liệu huấn luyện cho Qwen 2.5-VL

2.3.1. Loại dữ liệu

- Dữ liệu gồm 2 nhóm: chữ in và chữ viết tay, đa dạng font, bố cục, độ mờ và nhiễu.

2.3.2. Tiền xử lý và chuẩn hóa

- Chuẩn hóa ảnh, tăng tương phản, sửa nghiêng, lọc mẫu lỗi, chuẩn hóa văn bản ground truth (Unicode, khoảng trắng, loại ký tự không hợp lệ).

2.3.3. Định dạng dữ liệu theo phong cách hội thoại

Vì Qwen 2.5-VL được huấn luyện như một mô hình trò chuyện đa phương thức, dữ liệu được tổ chức dưới dạng các cặp **(ảnh, hội thoại)**. Ví dụ:

{
"image": "images/000123.png",
"conversations": [
{
"from": "human",
"value": "Hãy nhận dạng toàn bộ văn bản trong ảnh tài liệu sau và trả về nguyên văn."
},
{
"from": "gpt",
"value": "CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc
..."
}
]
}

2.4. Pipeline tinh chỉnh Qwen 2.5-VL bằng LoRA

2.4.1. Các bước tổng quát

Pipeline tinh chỉnh có thể được mô tả bằng mã giả (pseudocode) như sau:

```
# Bước 1: Chuẩn bị môi trường
load CUDA, cuDNN
install transformers, peft, qwen-vl, datasets
# Bước 2: Load mô hình gốc
base_model = load_qwen_vl_7b()
```

```
base_model.freeze_all_parameters()
# Bước 3: Cấu hình LoRA
lora_config = {
        target_modules:   ["q_proj",   "v_proj",   "k_proj",
"o_proj"],
    r: 16,
    alpha: 32,
    dropout: 0.05
}
model = apply_lora(base_model, lora_config)
# Bước 4: Load dataset
train_set = load_ocr_dataset("train.json")
val_set = load_ocr_dataset("val.json")

# Bước 5: Vòng lặp huấn luyện
for epoch in range(num_epochs):
    for batch in train_set:
        images, prompts, labels = batch
        logits = model(images, prompts)
        loss = cross_entropy_loss(logits, labels)
        loss.backward()
        optimizer.step()
        optimizer.zero_grad()
    evaluate(model, val_set)
# Bước 6: Lưu trọng số LoRA
save_lora_weights(model, "qwen_vl_ocr_lora")
```

2.4.2. Cấu hình huấn luyện cụ thể

2.5. Quy trình suy luận (inference) cho OCR

Sau khi tinh chỉnh, mô hình được dùng cho OCR theo pipeline sau:

1. **Nạp mô hình**: load Qwen 2.5-VL 7B và gắn trọng số LoRA đã huấn luyện.
2. **Tiền xử lý ảnh**: resize ảnh về kích thước phù hợp, chuẩn hóa pixel.
3. **Tạo prompt**: ví dụ:
    a. "Hãy nhận dạng toàn bộ văn bản trong ảnh sau"
4. **Gọi mô hình sinh văn bản**: dùng greedy decoding hoặc beam search với max_tokens đủ lớn.
5. **Hậu xử lý**:
    a. Loại bỏ token thừa (mô tả phụ, dấu ngoặc không cần thiết).
    b. Chuẩn hóa lại một số ký tự đặc biệt, khoảng trắng.

2.6. Thiết lập benchmark và tiêu chí đánh giá

Benchmark gồm: dữ liệu chữ in khó, chữ viết tay, và dữ liệu tổng hợp.

Các mô hình so sánh: PaddleOCR, Vintern, Qwen 2.5-VL trước và sau tinh chỉnh.

Tiêu chí đánh giá: CER và WER — hai độ đo quan trọng cho OCR tiếng Việt.

2.7. Kết luận chương

Chương 2 xây dựng phương pháp luận hoàn chỉnh: từ cơ sở lý thuyết VLM, kiến trúc Qwen 2.5-VL, kỹ thuật LoRA, thiết kế dữ liệu, pipeline huấn luyện–suy luận đến thiết lập benchmark.

Phương pháp đề xuất tận dụng ưu thế VLM để cải thiện OCR tiếng Việt trong điều kiện thực tế, đồng thời phù hợp triển khai trên GPU trung bình.

CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ

3.1. Bộ dữ liệu thực nghiệm

Chương 3 sử dụng ba bộ dữ liệu chính:

- UIT-HWDB: dữ liệu chữ viết tay tiếng Việt quy mô lớn (dạng dòng và đoạn), đa dạng nét chữ, kích thước ký tự và phong cách viết, giúp mô hình học được ngữ cảnh liên tục của văn bản viết tay.
- Vietnamese_original: dữ liệu chữ in đa dạng từ tài liệu hành chính, biểu mẫu, báo in, và ảnh chụp thực tế với nhiều điều kiện ánh sáng và chất lượng khác nhau.
- Vietnamese Handwriting OCR: dữ liệu chữ viết tay thu thập từ nhiều người, chất lượng ảnh rất khác nhau và có nhiều nhiễu thực tế; sau tiền xử lý còn 10.062 mẫu.

Ngoài ra, một benchmark độc lập gồm 500 ảnh (chữ in + viết tay) được xây dựng để đánh giá mô hình một cách khách quan.

3.2. Độ đo đánh giá (Metrics)

Hai metric chuẩn CER và WER được sử dụng:

- **CER** đo lỗi ký tự (thay thế / thiếu / thừa) – cực kỳ phù hợp với tiếng Việt do hệ thống dấu phức tạp.
- **WER** đo lỗi theo từ – quan trọng cho ứng dụng thực tế như KIE, trích xuất trường thông tin.

Cả hai metric đều minh bạch, dễ đối chiếu và phản ánh tốt hiệu suất OCR tiếng Việt.

3.3. Quy trình tinh chỉnh mô hình

Việc tinh chỉnh (SFT – Supervised Fine-Tuning) được thực hiện dựa trên Qwen 2.5-VL 7B, áp dụng phương pháp LoRA để giảm tải tài nguyên.

Tham số	**Giá trị**

LoRA rank	16
Alpha	32
Dropout	0.05
Learning rate	2e-5
Epoch	3
Batch size	1
Gradient accumulation	16
Max sequence length	3072
Scheduler	Cosine decay
Precision	BF16
GPU	RTX 3090 24GB

Bảng 3.1: Cấu hình huấn luyện

Mô hình Qwen 2.5-VL 7B được tinh chỉnh bằng kỹ thuật LoRA với rank=16,

alpha=32, dropout=0.05, learning rate 2e-5, 3 epoch.

Quy trình bao gồm:

- Khởi tạo mô hình Qwen-VL ở chế độ đóng băng trọng số gốc.
- Chèn LoRA vào các lớp attention (q/k/v/o proj).
- Chuẩn bị dữ liệu dạng JSONL theo format hội thoại.
- Huấn luyện bằng AdamW, BF16, gradient accumulation 16, max length 3072.
- Tối ưu bằng FlashAttention2, BitsAndBytes, Accelerate,… để vận hành trên GPU 24GB.
- Lưu lại module LoRA (~300MB) sau khi huấn luyện.

Mô hình hội tụ tốt, loss giảm đều theo epoch.

3.4. Kết quả thực nghiệm

Kết quả so sánh giữa các mô hình baseline và mô hình tinh chỉnh được thể hiện trong bảng

Mô hình	**CER**	**WER**
PaddleOCR [8]	16.80%	28.55%
Vintern [27]	3.12%	5.68%
Qwen 2.5-VL (pretrained) [17]	2.04%	4.92%
Qwen 2.5-VL +	**0.09%**	**2.02%**

LoRA		

Bảng 3.2: Kết quả thực nghiệm so sánh baseline

Kết quả cho thấy:

- Qwen 2.5-VL đã vượt PaddleOCR và Vintern ngay cả khi chưa tinh chỉnh.
- Sau tinh chỉnh, mô hình đạt **CER gần bằng 0**, cho thấy khả năng nhận dạng gần như hoàn hảo trên dữ liệu thử nghiệm.

Theo loại dữ liệu

Trong cả chữ in và chữ viết tay, Qwen + LoRA đều vượt trội:

- Printed: CER 0.04%
- Handwritten: CER 0.12%

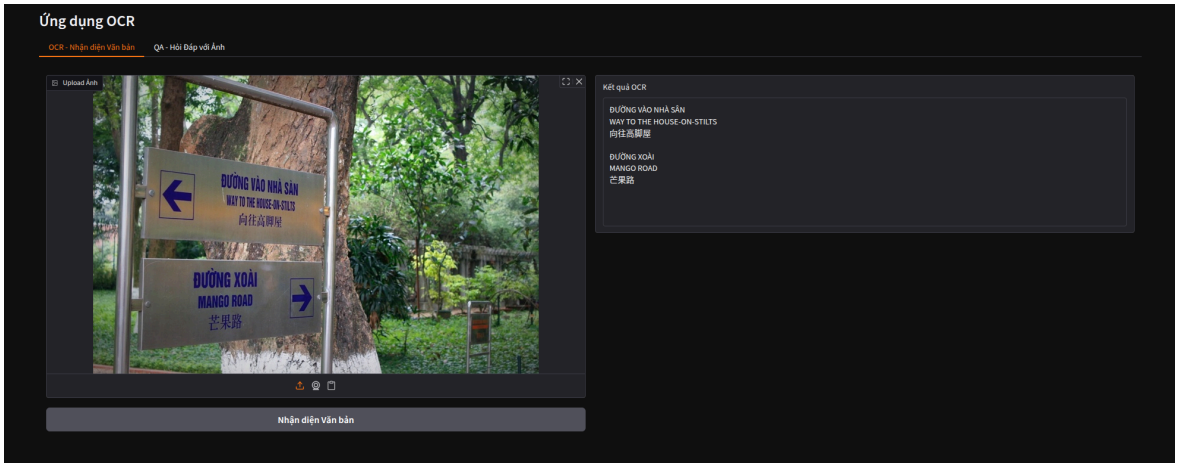
Theo độ dài chuỗi

Mặc dù lỗi tăng nhẹ với chuỗi dài (do tính chất sinh chuỗi autoregressive), CER vẫn <0.2% cho chuỗi >30 ký tự, chứng minh mô hình ổn định.

3.5. Xây dựng ứng dụng

Một ứng dụng web được xây dựng từ mô hình Qwen 3.5-VL đã finetune:

- Nhận dạng chữ in và viết tay.
- Hỏi đáp thông tin từ ảnh (Document-QA).
- Thử nghiệm cho thấy nhận dạng chữ in rất chính xác; chữ viết tay khó vẫn còn hạn chế, nhưng mô hình hiểu ngữ cảnh và trích xuất trường thông tin hiệu quả.



Hình 3.1: Thử nghiệm với hình ảnh đa ngôn ngữ

3.7. Kết luận chương

Chương này đã chứng minh rằng việc tinh chỉnh mô hình Qwen 2.5-VL 7B bằng LoRA trên bộ dữ liệu tiếng Việt giúp mô hình đạt chất lượng OCR vượt trội, đặc biệt trên tài liệu khó và chữ viết tay. Các chỉ số CER và WER đều giảm mạnh, mô hình hội tụ tốt và ổn định, đồng thời vượt xa các baseline truyền thống như PaddleOCR. Kết quả này khẳng định hướng tiếp cận dùng VLM cho OCR tiếng Việt là hoàn toàn khả thi và có tiềm năng ứng dụng rộng rãi trong các hệ thống số hóa tài liệu.

KẾT LUẬN

Đề án đã nghiên cứu và triển khai một hướng tiếp cận mới cho bài toán nhận dạng và trích xuất thông tin từ ảnh tài liệu tiếng Việt dựa trên mô hình thị giác–ngôn ngữ lớn Qwen 2.5-VL. Bằng cách tinh chỉnh mô hình thông qua kỹ thuật LoRA trên các bộ dữ liệu chữ in và chữ viết tay được chuẩn hóa, mô hình đã thể hiện khả năng học đặc thù tiếng Việt một cách hiệu quả. Các kết quả thực nghiệm cho thấy mô hình đạt độ chính xác vượt trội so với các baseline truyền thống như PaddleOCR và Vintern, đặc biệt ở các trường hợp văn bản khó, nhiễu hoặc viết tay.

Ngoài ra, mô hình sau khi tinh chỉnh còn cho thấy tính ứng dụng cao thông qua hệ thống demo có khả năng nhận dạng văn bản và trích xuất thông tin trực tiếp từ ảnh tài liệu. Điều này mở ra tiềm năng lớn cho việc triển khai thực tế trong các hệ thống số hóa dữ liệu, tự động hóa quy trình nhập liệu và các ứng dụng VQA dựa trên tài liệu tiếng Việt.

Mặc dù đạt được những kết quả khả quan, đề án vẫn còn một số hạn chế, như mô hình đôi khi chưa xử lý tốt các kiểu chữ viết tay quá phức tạp. Do vậy, các hướng phát triển tiếp theo có thể bao gồm mở rộng quy mô và đa dạng dữ liệu huấn luyện, kết hợp thêm các kỹ thuật tăng cường dữ liệu, và thử nghiệm các phiên bản VLM lớn hơn để cải thiện độ chính xác trên dữ liệu thực tế.

Nhìn chung, đề án đã chứng minh tính khả thi và hiệu quả của việc ứng dụng mô hình thị giác–ngôn ngữ lớn cho OCR tiếng Việt, đồng thời cung cấp một hướng tiếp cận hiện đại, tiềm năng cho các bài toán phân tích và xử lý tài liệu trong bối cảnh chuyển đổi số hiện nay.