Trường Đại học Khoa học tự nhiên

Khoa Công nghệ Thông tin

CSC15006 – Nhập môn Xử lý ngôn ngữ tự nhiên

BÁO CÁO

FINE-TUNING DEEPSEEK-OCR WITH VIETNAMESE DATASET

Lớp 23_22

Trương Bảo Thiên Ân - 23120019

Mục lục

Danh mục hình ảnh

1 Đặt vấn đề

1.1 OCR tiếng Việt

Sự bùng nổ của kỷ nguyên số đã đặt ra nhu cầu cấp thiết về việc chuyển đổi hàng tỷ trang tài liệu giấy sang định dạng số có thể chỉnh sửa và tìm kiếm được. Trong lĩnh vực OCR, bài toán nhận dạng chữ viết tay luôn được coi là một trong những thách thức lớn nhất. Khác với văn bản in máy có cấu trúc đồng nhất, chữ viết tay mang tính cá nhân cao, thay đổi về nét chữ, độ nghiêng, và mật độ ký tự.

Đối với tiếng Việt, thách thức này trở nên lớn hơn do đặc thù ngôn ngữ. Tiếng Việt sử dụng hệ chữ cái Latinh nhưng được bổ sung bởi một hệ thống thanh dấu phức tạp. Sự hiện diện của các dấu như sắc, huyền, hỏi, ngã, nặng, cùng với các dấu mũ (â, ê, ô) tạo ra một không gian ký tự dày đặc. Trong chữ viết tay, các dấu này thường xuyên bị viết nhanh, dính liền vào ký tự hiện tại hoặc bị đặt lệch vị trí, làm cho các mô hình OCR nhận dạng sai.

1.2 Deepseek OCR

DeepSeek-OCR là hệ thống nén thông tin đa phương thức tiên tiến hoạt động trên cơ chế Encoder-Decoder, nổi bật với kỹ thuật Contextual Optical Compression. Cốt lõi là kiến trúc DeepEncoder (380M tham số) đảm nhiệm việc xử lý hình ảnh qua ba giai đoạn tinh vi: sử dụng SAM-base để nhận diện chi tiết nhỏ (như dấu thanh), nén dữ liệu thị giác xuống 16 lần để loại bỏ nhiễu, và dùng CLIP-large để thấu hiểu bố cục toàn cục. Chuỗi token thị giác cô đọng này sau đó được chuyển đến bộ giải mã DeepSeek-3B-MoE - một mô hình ngôn ngữ lớn sử dụng kiến trúc "Mixture-of-Experts" linh hoạt - giúp hệ thống suy luận nhanh chóng, tự động sửa lỗi dựa trên ngữ cảnh và sinh ra văn bản chính xác mà không gây bùng nổ tài nguyên tính toán.

2 Phương pháp nghiên cứu

2.1 Thu thập dữ liệu

Để tiến hành fine-tune model, tác giả đã sử dụng bộ dữ liệu **UIT-HWDB**

2.2 Quá trình tiền xử lý

1. **Chuẩn hóa văn bản:** Toàn bộ nhãn được chuyển về định dạng Unicode NFC dựng sẵn để đảm bảo tính nhất quán. Các ký tự không cần thiết được loại bỏ.

2. **Định dạng Huấn luyện:** Dữ liệu được chuyển đổi sang định dạng JSON cho phù hợp với định dạng của Unsloth. Mỗi mẫu dữ liệu bao gồm đường dẫn ảnh và chuỗi văn bản tương ứng. Sử dụng đường dẫn ảnh thay vì định dạng PIL nhằm giảm áp lực phải load toàn bộ dataset lên RAM.

   Ví dụ:

```
{
  "messages": [
    {
      "role": "<|User|>",
```

```
      "content": "<image>\nFree OCR. ",
      "images": ["UIT_HWDB_word/train_data/29/114.jpg"]
    },
    {
      "role": "<|Assistant|>",
      "content": "hợp"
    }
  ]
}
```

3. **Tạo dataset để train và test:** Để phù hợp với tài nguyên giới hạn (Google Colab T4), tác giả không lấy toàn bộ dataset để train mà chỉ lấy một phần:

   Đối với tập train:

   - Sử dụng 50% dữ liệu từ tập Line và Paragraph để đảm bảo mô hình học được ngữ cảnh câu.
   - Sử dụng một tập con 500 mẫu từ tập Word để tăng cường khả năng nhận diện từ đơn lẻ.

   Đối với tập test:

   - Sử dụng 100% dữ liệu từ tập Line, tập con 20 mẫu từ tập Paragraph, tập con 200 mẫu từ tập Word để đánh giá mô hình.

4. **Phân chia dữ liệu:** Dataset **UIT-HWDB** đã chia sẵn thành tập train và tập test, đảm bảo không có sự chồng lặp về người viết giữa các tập để đánh giá chính xác khả năng tổng quát hóa.

2.3 Chi tiết cơ chế Fine-tune

Chiến lược: QLoRA với cơ chế PERT (Parameter-Efficient Fine-Tuning)

Module mục tiêu: Các adapter được gắn vào các lớp projection `q_proj`, `k_proj`, `v_proj`, `o_proj`) và các lớp MLP `gate_proj`, `up_proj`, `down_proj`

Hyperparameters:

- **QLoRA rank (r):** 16
- **Alpha:** 16
- **Learning rate:** 2e-4
- **Batch size:** 2 kết hợp với **Gradient Accumulation Steps** là 4, giúp gradient ổn định hơn mà không gây tràn VRAM.
- **Optimizer:**: adamw_8bit

- **Scheduler::** linear
- **Epochs::** 1 epoch cho toàn bộ training dataset

3 Thiết kế thực nghiệm

3.1 Môi trường chạy

Chương trình được chạy trên nền tảng Google Colab với phần cứng là T4 16Gb VRAM

3.2 Metric đánh giá

Sử dụng metric CER (Character Error Rate) để đo lường sự sai khác giữa văn bản được mô hình dự đoán và nhãn thực tế dựa trên từng ký tự.

3.3 Thống kê dữ liệu huấn luyện và kiểm thử

Bảng 1 trình bày chi tiết số lượng mẫu được sử dụng từ các bộ dữ liệu UIT_HWDB khác nhau để tạo thành tập huấn luyện (train.json) và tập kiểm thử (test.json).

Table 1: Thống kê số lượng mẫu trong tập Train và Test

Tập dữ liệu	**Train (Sử dụng / Tổng)**	**Test (Sử dụng / Tổng)**
UIT_HWDB_line	3514 / 7028	201 / 201
UIT_HWDB_paragraph	1113 / 1113	20 / 31
UIT_HWDB_word	1000 / 107607	200 / 2881
Tổng cộng	**5627**	**421**

3.4 Kịch bản đánh giá

- Baseline: Mô hình DeepSeek-OCR gốc chưa qua tinh chỉnh. Đánh giá khả năng nhận dạng Tiếng Việt của mô hình.
- Fine-tuned Model: Mô hình sau khi đã được huấn luyện trên tập train.json

4 Kết quả

4.1 Kết quả CER

Hiệu quả của quá trình fine-tune được đánh giá dựa trên CER. Kết quả cho thấy sự cải thiện vượt bậc sau khi fine-tune , giảm thiểu đáng kể lỗi ở cả cấp độ từ và cấp độ dòng.

Bảng 2 tóm tắt kết quả so sánh trước và sau khi huấn luyện:

Table 2: So sánh chỉ số CER trước và sau khi fine-tuning

Cấp độ đánh giá	**Basline**	**Fine-tuned**	**Cải thiện (Điểm)**
Word	7.6715	**0.5789**	+7.0926
Line	1.4304	**0.2574**	+1.1730

4.2 Phân tích

Trước khi tinh chỉnh, mô hình gốc thường xuyên gặp hiện tượng hallucination, dự đoán sai lệch hoàn toàn sang các công thức toán học (ví dụ: $4\pi^2$) hoặc tiếng Anh vô nghĩa (ví dụ: "him angry") đối với dữ liệu chữ viết tay tiếng Việt.

Sau khi tinh chỉnh, mô hình đã học được cấu trúc ngôn ngữ tiếng Việt và hình dạng ký tự viết tay, cho ra kết quả sát với nhãn thực tế hơn rất nhiều.

4.2.1 Phân tích mức độ Word

Table 3: So sánh kết quả mức độ Word

Input Image	Ground Truth	Baseline	Fine-tuned
	đốn	[# an]	đơn
	hỏi	[4 π^2]	hán

4.2.2 Phân tích mức độ Line

Dưới đây là kết quả chi tiết trên một số mẫu dữ liệu dòng, so sánh giữa nhãn thực tế, mô hình trước và sau khi tinh chỉnh.

Figure 1: Mẫu kiểm thử số 1 (Line)

Table 4: Chi tiết kết quả nhận diện của mẫu 1

Ground truth	hiệu hiệp sĩ.
Baseline	him angry
Fine-tuned	hình như là

Figure 2: Mẫu kiểm thử số 2 (Line Level)

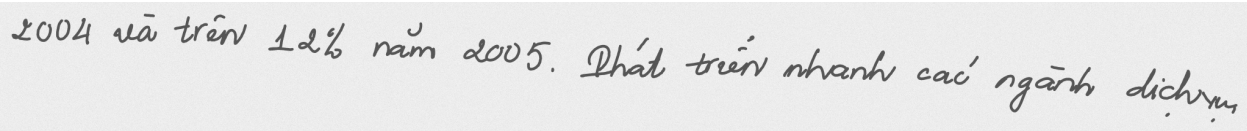

Table 5: Chi tiết kết quả nhận diện của mẫu 2

Ground truth	2004 và trên 12% năm 2005. Phát triển nhanh các ngành dịch vụ,
Baseline	2004 năm triển lãi và năm 2005. Đặt triển nhạnh các ngành dịch vụ.
Fine-tuned	2004 của triển lãi và năm 2005. Đất triển nhạnh các ngành dịch vụ

4.2.3 Phân tích mức độ Paragraph

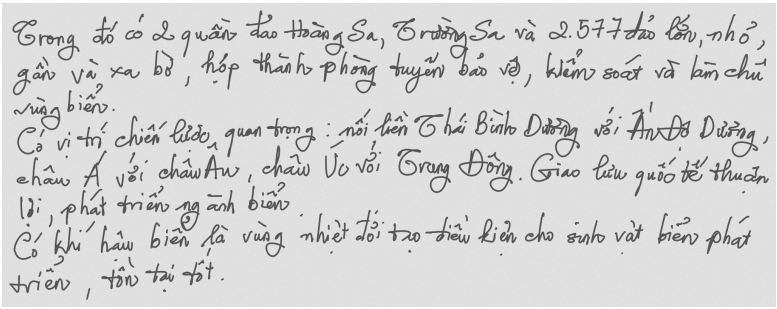

Figure 3: Mẫu Paragraph số 1

Table 6: Chi tiết kết quả mẫu Paragraph 1

Ground truth	Trong đó có 2 quần đảo Hoàng Sa, Trường Sa và 2.577 đảo lớn, nhỏ, gần và xa bờ, hợp thành phòng tuyến bảo vệ, kiểm soát và làm chủ vùng biển. Có vị trí chiến lược quan trọng : nối liền Thái Bình Dương với Ấn Độ Dương, châu Á với châu Âu, châu Úc với Trung Đông. Giao lưu quốc tế thuận lợi, phát triển ngành biển. Có khí hậu biển là vùng nhiệt đới tạo điều kiện cho sinh vật biển phát triển, tồn tại tốt.
Baseline	Grong đó có 2 quần tảo hoàng Sa, Trường Sa và 2.577 tảo bản, nhỏ, gần và xa bỏ, hóp thành phong tuyên báo về kiêm soát và bàn chủ rừng biển. Có vị thứ chiến lược quan trọng: mọi liên Thái Bình Dương và Anh Dương, cháu Anh Anh, cháu Võ Võ, Trung Đông. Giao lưu quốc tế thuận lợi, phát triển ngành bản. Có khi hai biên la rừng nhất đối tạo điều kiện cho sinh vật biên phát triển, tôn tài tốt.
Fine-tuned	Trong đó có 2 quần đảo hoàng Sa, Trường Sa và 2.577 đảo lớn, nhỏ, gần và xa bờ, hồp thành phong tuyến bảo vệ, kiểm soát và bàn chủ rừng biển. Có vị thứ chiến lược quan trọng : nói liên Chí Bình Dương và Anh Dương, cháu A với cháu Anh, cháu Võ Võ Trung Đông. Giao lưu quốc tế thuận lợi, phát triển ngành biển. Có khi hai biển là vùng nhiệt độ tạo điều kiện cho sinh vật biển phát triển, tôn tài tốt.
Nhận xét	Kết quả dự đoán gần như trùng khớp hoàn toàn

Khác với mức độ Word và Line, việc nhận diện ở mức độ Paragraph đặt ra thách thức lớn hơn về khả năng ghi nhớ chuỗi dài.

Trường hợp nhận diện tốt: Với các đoạn văn bản có cấu trúc rõ ràng, mô hình Fine-tuned thể hiện khả năng nắm bắt ngữ cảnh tốt, nhận diện chính xác cả các dấu câu và chữ số phức tạp.

Các trường hợp nhận diện tệ, làm CER trung bình tăng: Mặc dù phần lớn các mẫu đều có kết quả tốt, chỉ số CER trung bình trên tập Paragraph lại tăng (từ 3.9 lên 5.0). Nguyên nhân chính đến từ hiện tượng **"Lặp từ vô tận"** ở một số ít mẫu ngoại lai.

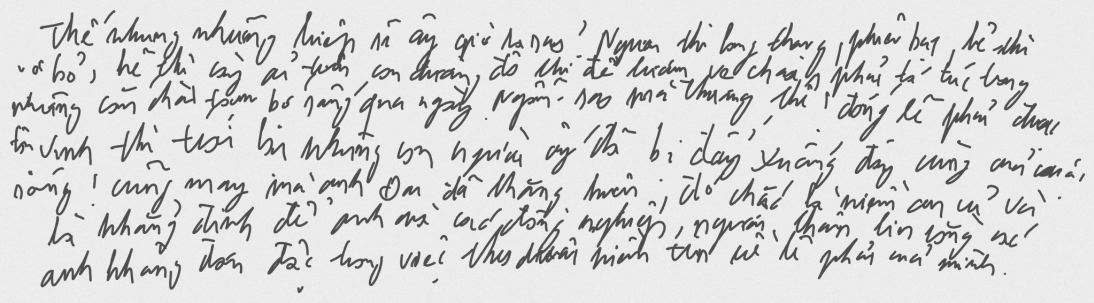

Figure 4: Mẫu Paragraph số 2 (Mẫu gây lỗi Outlier)

Table 7: Chi tiết lỗi lặp từ ở mẫu Paragraph 2

Ground truth	Thế nhưng những hiệp sĩ ấy giờ ra sao? Người thì lang thang, phiêu bạt, kẻ thì vợ bỏ, kẻ thì cày ải trên những con đường đô thị để lượm ve chai, phải tá túc trong những căn chòi tạm bợ sống qua ngày. Ngẫm sao mà thương thế! Đáng lẽ phải được tôn vinh thì trái lại những con người ấy đã bị đẩy xuống đáy cùng của cuộc sống! Cũng may mà anh Đại đã thắng kiện, đó chắc là một niềm an ủi và là một khẳng định để anh và các đồng nghiệp, người thân tin rằng các anh không đơn độc trong việc theo đuổi niềm tin về lẽ phải của mình.
Baseline	The charming, charming, charming, in an airy, airy, airy, airy, airy, airy, airy, air, airy, airy, airy, airy, airy, airy, in an airy, airy, airy, airy, airy, air, airy. The charming, charming, charming, in an airy, airy, airy, airy. The charming, charming, charming, in an airy, airy, in an airy,...
Fine-tuned	The chính mình không biết nó gì là sao? Ngược thì bên trong chính bà chủ nói đó, không có chuyện để chuyển phích ra bà có không biết còn chứa tâm bỏ ngã ngã ngã ngã. Sẽ mở chúng thì tôi có thể không biết chính bà chủ không biết chính bà chủ không biết chính bà chủ không bị chính bà chủ không biết chính bà chủ không biết chúng. Tôi có thể không biết chính bà chủ không biết chính bà không biết chính bà không biết chính bà không biết chính bà không biết chính không biết chính không biết chính không biết chính không biết chính không biết chúng không biết chính không biết chính không biết chính không biết chứa tâm bỏ ngã ngã ngã ngã...
Giải thích	Mô hình gặp lỗi trong việc xác định token kết thúc câu (EOS), dẫn đến việc sinh ra hàng nghìn ký tự vô nghĩa. Vì công thức CER tính dựa trên số lượng ký tự sai, một mẫu lỗi này có thể tạo ra CER $> 1000\%$, kéo lùi chỉ số trung bình của toàn bộ tập dữ liệu.

5 Kết luận

Kết quả cho thấy sự thay đổi rõ rệt của mô hình sau quá trình fine-tuning. Mô hình Baseline ban đầu hoàn toàn không phù hợp với tác vụ nhận diện chữ viết tay tiếng Việt (CER > 7.0). Tuy nhiên, sau khi được tinh chỉnh, CER đã giảm mạnh xuống mức chấp nhận được (dao động từ 0.25 đến 0.57 ở mức độ Word và Line).

Mặc dù mô hình đạt được kết quả khả quan, mô hình vẫn còn một số hạn chế:

1. **Hiệu năng ở Word và Line:** Mô hình thể hiện khả năng học tốt các đặc trưng hình thái của chữ viết tay tiếng Việt. Tuy nhiên, vẫn tồn tại các sai sót nhỏ liên quan đến:

- Nhận diện dấu câu.
- Sự nhầm lẫn giữa các từ có hình dạng ký tự tương đồng, ví dụ: "đốn" → "đơn", "hỏi" → "hán". Điều này cho thấy mô hình đôi khi vẫn dựa quá nhiều vào đặc trưng hình ảnh cục bộ mà chưa tận dụng hết ngữ cảnh.

2. **Vấn đề lặp từ ở Paragraph:** Chỉ số CER trung bình ở tập Paragraph tăng lên mức 5.0, nhưng phân tích cho thấy đây **không phải do sự suy giảm năng lực nhận diện**. Nguyên nhân chính đến từ hiện tượng "lặp từ vô tận" ở một số ít mẫu đặc biệt, chữ xấu, khiến độ dài văn bản dự đoán tăng đột biến và làm sai lệch chỉ số trung bình. Nếu loại bỏ các mẫu này, hiệu năng thực tế trên các đoạn văn bản dài vẫn đảm bảo độ chính xác cao.

3. **Hướng cải thiện và tối ưu hóa:** Để khắc phục các hạn chế trên, các giải pháp tiềm năng bao gồm:

   - Áp dụng các chiến lược post-processing như ‘repetition_penalty‘ hoặc giới hạn ‘no_repeat_ngram_size‘ trong quá trình inference để loại bỏ lỗi lặp từ.
   - Tăng cường dữ liệu huấn luyện tập trung vào các cặp từ dễ gây nhầm lẫn để cải thiện khả năng phân biệt.
   - Mở rộng tập train để có thể học được nhiều hơn.
   - Áp dụng Synthetic Data để tạo ra nhiều data training hơn, giúp model học được các đặc trưng tinh tế hơn của chữ viết tay Tiếng Việt

6 Source Code

Source Code này bao gồm toàn bộ pipeline xử lý, và cả các cell để đánh giá mô hình.