ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

Báo cáo: Fine-tune Deepseek-OCR với dữ liệu Tiếng Việt

Môn học: Nhập Môn Xử Lý Ngôn Ngữ Tự Nhiên

Sinh viên thực hiện:

Đinh Xuân Khương - 23127398

Giảng viên hướng dẫn:

Đinh Điền

Nguyễn Hồng Bửu Long

Lương An Vinh

Ngày 11 tháng 12 năm 2025

Mục lục

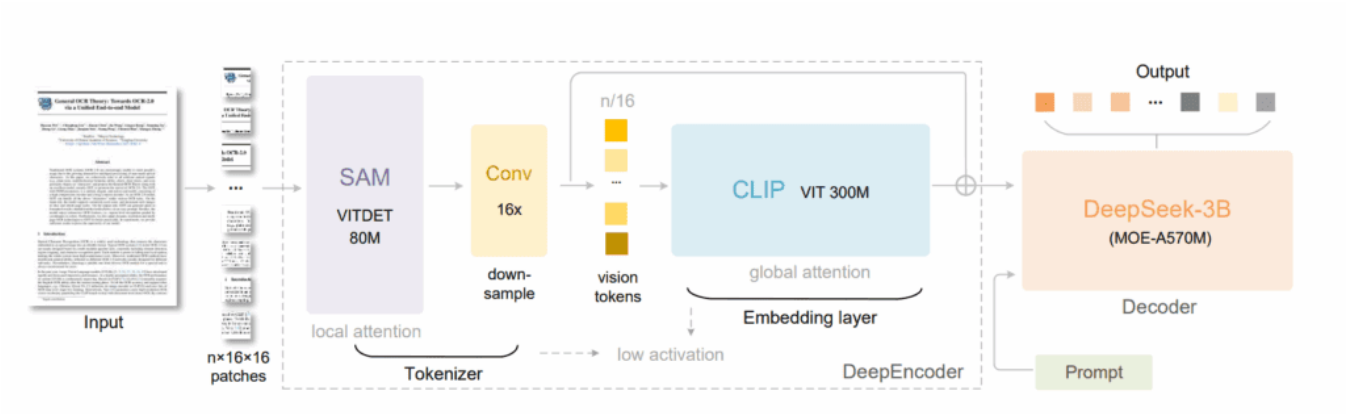
1 Bối cảnh hiện nay

1.1 Giới thiệu về OCR

Nhận dạng ký tự quang học (Optical Character Recognition - OCR) là kỹ thuật chuyển đổi hình ảnh chứa văn bản thành dạng văn bản số có thể xử lý được. OCR giữ vai trò quan trọng trong số hóa tài liệu, trích xuất thông tin và tự động hóa nhiều quy trình. Tuy nhiên, OCR cho tiếng Việt gặp nhiều thách thức do đặc thù ngôn ngữ, bao gồm:

- Hệ thống dấu phức tạp với nhiều kiểu dấu kết hợp.
- Chữ cái có hình dạng tương tự nhau dễ gây nhầm lẫn cho mô hình.
- Tài liệu thực tế có thể bị nhiễu, mờ, nghiêng hoặc có chất lượng kém.
- Sự đa dạng về font chữ, kích thước và bố cục văn bản.
- Chữ viết tay không được rõ ràng và đẹp.

1.2 Tổng quan về mô hình DeepSeek-OCR



Hình 1: Kiến trúc Deepseek-OCR

Kiến trúc của DeepSeek-OCR bao gồm ba thành phần chính: *Tokenizer*, *DeepEncoder* và *Decoder*. Toàn bộ hệ thống được thiết kế theo hướng kết hợp giữa mô hình thị giác (Vision Transformer) và mô hình ngôn ngữ lớn (LLM) nhằm tối ưu cho tác vụ nhận dạng văn bản trong ảnh.

1. Tokenizer

Ảnh đầu vào được chia thành các patch kích thước 16×16. Các patch này được đưa vào module **SAM/VITDET** để trích xuất đặc trưng cục bộ. Sau đó, một tầng **Conv 16×16** thực hiện giảm mẫu (downsampling) và chuyển đổi ảnh thành các vision tokens. Tokenizer có nhiệm vụ:

- Biến đổi ảnh thô thành chuỗi token đặc trưng.
- Giữ lại thông tin không gian quan trọng phục vụ nhận dạng ký tự.

2. DeepEncoder

Chuỗi vision tokens từ Tokenizer được đưa vào **CLIP-ViT (300M)** để mô hình hóa đặc trưng toàn cục. DeepEncoder đóng vai trò:

- Kết hợp thông tin cục bộ và toàn cục từ ảnh.
- Tạo ra embedding có ý nghĩa ngữ cảnh, phù hợp cho mô hình ngôn ngữ xử lý.
- Tối ưu hóa thông tin bằng cơ chế *low activation*, giúp giảm độ phức tạp tính toán.

3. Decoder (DeepSeek-3B-MoE)

Bộ giải mã sử dụng mô hình ngôn ngữ DeepSeek-3B (MOE-A570M). Decoder nhận embedding từ DeepEncoder cùng với *prompt* OCR, sau đó sinh ra chuỗi ký tự đầu ra. Chức năng chính:

- Dự đoán văn bản theo từng token dựa vào thông tin hình ảnh được mã hóa.
- Áp dụng cơ chế attention giữa embedding ảnh và quá trình sinh văn bản.

Có thể nói một cách dễ hiểu, DeepSeek-OCR là mô hình kết hợp mạnh mẽ giữa encoder thị giác nhiều tầng và decoder ngôn ngữ lớn, cho phép nhận dạng văn bản chính xác trong nhiều điều kiện hình ảnh phức tạp.
Chúng ta có thể đọc rõ hơn thông qua paper này: DeepSeek-OCR: Contexts Optical Compression.

2 Phương pháp

2.1 Dữ liệu huấn luyện

Dữ liệu sử dụng là bộ dữ liệu: UIT-HWDB dataset.
Với giấy phép có thể xem ở đây: UIT-HWDB.

2.2 Môi trường huấn luyện

Mô hình được huấn luyện trên môi trường GPU P100 của Kaggle với các thông số:

- **GPU:** 1xTesla P100 , having 3584 CUDA cores, 16GB(16.28GB Usable) GDDR6 VRAM
- **CPU:** Intel(R) Xeon(R) CPU @ 2.00GHz , 39MB Cache
- **RAM:** 29 GB Available
- **Disk:** 57.6 GB Available

2.3 Chi tiết về Fine-tune

Mô hình được fine-tune theo chiến lược "fine-tune và đánh giá" dựa theo hướng dẫn của: Unsloth. Với các thông số được tinh chỉnh:

Tham số	**Giá trị**
Tokenizer	FastVisionModel deepseek_ocr
Image size	640px
Base size	1024px
Batch size mỗi thiết bị	2
Gradient accumulation steps	4
Warmup steps	5
Max steps	100
Learning rate	2.10^{-4}
Logging steps	1
Optimizer	adamw_8bit
Weight decay	0.001
Scheduler	Linear
Seed	3407
FP16	Sử dụng nếu BF16 không hỗ trợ
BF16	Sử dụng nếu được hỗ trợ
Output directory	outputs
Báo cáo kết quả (report_to)	none
Số luồng DataLoader	2
remove_unused_columns	False (bắt buộc cho vision finetuning)

Bảng 1: Các siêu tham số sử dụng trong quá trình huấn luyện.

2.4 Vấn đề nảy sinh và hậu xử lý

Vì chức năng **infer** của mô hình chỉ in ra màn hình và trả về kiểu dữ liệu **NoneType** thế nên tôi không thể lấy được dữ liệu đầu ra của mô hình bằng cách gán vào biến kết quả.

Hướng xử lý được thầy Bửu Long gợi ý: Redirect Output, thay vì in ra màn hình, tôi có thể ghi vào một file, từ đó có được kết quả để đánh giá mô hình.

Ngoài ra, sau khi lấy được dữ liệu thô bằng cách Redirect Output, có một vài vấn để nhỏ cần được xử lý, cụ thể như sau:

- Output chứa những dòng thông tin không cần thiết (BASE, PATCHES, ===).

```
====================
BASE:  torch.Size([1, 256, 1280])
PATCHES:  torch.Size([9, 100, 1280])
====================
Vui tôi tiếp vấn muốn tham gia Xuân Đại. Ngày tôi thôi tự bảo "cô phẩm của"
y_true: Viết tiếp về vụ anh Trương Xuân Đại. Ngay từ khi tờ báo có bài " Số phận của
====================
BASE:  torch.Size([1, 256, 1280])
PATCHES:  torch.Size([9, 100, 1280])
====================
making no doubt lei is that that's what's in, there's been a lot of big to us this set.
y_true: những người nói lên sự thật " phát hành, nhiều bạn đọc đã bày tỏ sự chia sẻ,
====================
```

Hình 2: Dữ liệu đầu ra của mô hình cơ sở

- Dữ liệu dự đoán của đoạn văn (paragraph) không nằm trên cùng một hàng.

```
====================
BASE:  torch.Size([1, 256, 1280])
PATCHES:  torch.Size([3, 100, 1280])
====================
Tiếp tục nhận rộng, các mỏ hình huy dòng vốn đâù tư. Tập trung khảo
thuật các nguồn vốn, cho ngân sách. Tổng kết quả trình có phản hại
doanh nghiệp nhà nước và việc tổ chức tập xếp lại doanh nghiệp nhà
nước. Trên khai quyết định của TP về một số chính sách vừa đâù
tư vốn các dự án có vốn đâù tư nước ngoài. Ra sát lại, chúng
tự hội hòa, phát triển xã hội học tập, vừa vận hòa - xã hội: đẩy mạnh
cách cho giao dục - đào tạo.
y_true: Tiếp tục nhân rộng các mô hình huy động vốn đâù tư. Tập trung khai thác các
====================
BASE:  torch.Size([1, 256, 1280])
PATCHES:  torch.Size([5, 100, 1280])
====================
Nàng cao này là quản lý nhà nước trên các lĩnh vực kinh doanh dịch vụ văn hóa; thời
y_true: Nâng cao năng lực quản lý nhà nước trên các lĩnh vực kinh doanh dịch vụ văn
====================
```

Hình 3: Dữ liệu đầu ra của mô hình đã tinh chỉnh

Hướng giải quyết: Chỉ cần loại bỏ những dòng không cần thiết và ghép những dòng của đoạn văn thành một dòng duy nhất. Code hậu xử lý này đã được đính kèm cùng bài nộp.

3 Thực nghiệm

3.1 Các mô hình

Trong dữ liệu của UIT-HWDB đã chia sẵn các tập huấn luyện và tập kiểm thử, thế nên tôi không cần chia nữa.
Cụ thể, vì giới hạn của chi phí tính toán nên tôi huấn luyện 2 mô hình:

- Mô hình huấn luyện với **80 279** ảnh.
- Mô hình huấn luyện với **100 743** ảnh.

Và kiểm tra trên tập huấn luyện với **3 113** ảnh.
Người đọc/thầy có thể xem chi tiết quy trình huấn luyện của mô hình được huấn luyện với hơn 100000 ảnh sau hơn 7 tiếng ở đây: Kaggle Notebook.

3.2 Các chỉ số đánh giá

Trong bài toán nhận dạng ký tự quang học (OCR), việc đánh giá chất lượng mô hình cần dựa trên các thước đo phản ánh được độ chính xác ở cả mức ký tự và mức từ. Do đó, ba chỉ số **Character Error Rate**, **Word Error Rate** và **Exact Match** được lựa chọn vì các lý do sau:

- **Character Error Rate (CER)**: Là thước đo dựa trên khoảng cách Levenshtein giữa chuỗi dự đoán và chuỗi mục tiêu ở mức ký tự. CER phản ánh số lỗi *thêm, xoá, hoặc thay thế* ký tự so với tổng số ký tự trong văn bản. CER đặc biệt quan trọng đối với tiếng Việt vì hệ thống dấu phức tạp, mô hình dễ mắc lỗi ở từng ký tự riêng lẻ.
- **Word Error Rate (WER)**: Được tính dựa trên lỗi ở mức từ. WER giúp đánh giá khả năng mô hình giữ đúng cấu trúc và nghĩa của câu. Đây là thước đo phù hợp khi ứng dụng OCR trong các tài liệu dài, có tính liền mạch về ngữ nghĩa. WER bổ sung cho CER bằng cách phản ánh chất lượng nhận dạng ở cấp độ cao hơn.

- **Exact Match (EM)**: Là tỷ lệ số mẫu mà mô hình dự đoán *chính xác hoàn toàn* so với ground truth. EM là chỉ số nghiêm ngặt, giúp đánh giá khả năng mô hình tái tạo văn bản không sai sót. Chỉ số này quan trọng trong các ứng dụng đòi hỏi độ chính xác tuyệt đối như trích xuất thông tin hành chính, mã số, biểu mẫu, hoặc tài liệu pháp lý.

Việc kết hợp cả ba chỉ số giúp đánh giá mô hình một cách toàn diện: *CER* đánh giá độ chính xác chi tiết, *WER* đánh giá độ chính xác ngữ cảnh, và *EM* đánh giá độ chính xác tổng thể.

4 Kết quả và đánh giá

4.1 Chỉ số toàn bộ mô hình

Metrics	**Base Model**	**80k Images**	**100k Images**
Overall CER	1.5697	**0.5190**	1.1570
Overall WER	1.6488	**0.7488**	1.4708
Overall Exact Match (EM)	0.0074	0.2547	**0.2560**
Mean CER	5.7690	**0.4760**	0.4928
Mean WER	2.6033	**0.7247**	0.7424
Single Word CER	5.1561	0.4634	**0.4578**
Single Word WER	2.7108	0.7372	**0.7330**
Single Word EM	0.0080	0.2743	**0.2764**
Sequence CER	0.6644	**0.5330**	1.4166
Sequence WER	1.1640	**0.7522**	1.8249
Sequence EM	0.0000	**0.0087**	0.0000

Bảng 2: So sánh hiệu suất giữa mô hình gốc và các mô hình được huấn luyện với 80k và 100k ảnh.

Trong đó:

- Overall: Các số liệu được đánh giá trên toàn bộ tập kiểm tra sử dụng thư viện **jiwer**.
- Mean: Các số liệu được lấy trung bình lỗi trên từng mẫu dữ liệu của toàn bộ tập kiểm tra.
- Single Word: Số liệu được đánh giá trên dữ liệu chỉ gồm một chữ duy nhất.
- Sequence: Số liệu được đánh giá trên dữ liệu gồm câu và đoạn văn.

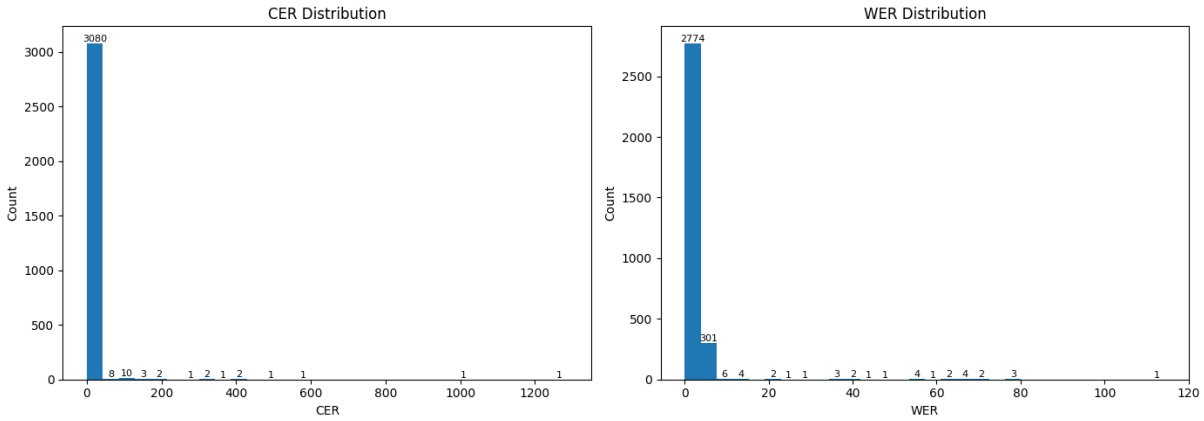
Nhận xét kết quả

Dựa trên bảng số liệu, có thể đưa ra một số nhận định quan trọng như sau:

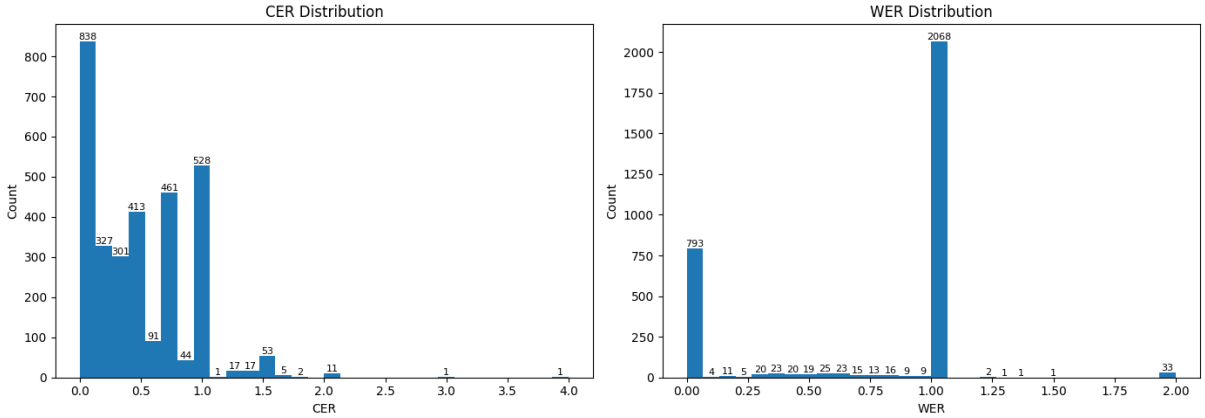
- **Hiệu suất của mô hình huấn luyện với 80k ảnh vượt trội so với cả mô hình gốc và mô hình được huấn luyện với 100k ảnh ở hầu hết các chỉ số**. Điều này thể hiện rõ qua các thước đo tổng hợp như Overall CER (0.5190) và Overall WER (0.7488), đều thấp hơn đáng kể so với mô hình gốc và 100k ảnh.
- **Độ chính xác tổng thể (Overall EM)** của mô hình 80k ảnh (0.2547) tăng mạnh so với mô hình gốc (0.0074), đồng thời chỉ kém nhẹ mô hình 100k ảnh (0.2560). Điều này cho thấy việc tăng dữ liệu từ 80k lên 100k ảnh không mang lại cải thiện đáng kể về EM.
- **Ở nhóm đánh giá mức ký tự trung bình (Mean CER) và mức từ trung bình (Mean WER)**, mô hình 80k ảnh tiếp tục đạt kết quả tốt nhất. Mean CER giảm từ 5.7690 xuống 4.4760, và Mean WER giảm từ 2.6033 xuống 0.7247 — mức cải thiện rất lớn.
- **Trong tác vụ nhận dạng từng từ đơn lẻ (Single Word)**, mô hình 100k ảnh thể hiện tốt nhất với Single Word CER = 0.4578 và Single Word WER = 0.7330. Có thể lý do chính là 20k dòng dữ liệu thêm vào chỉ gồm các Single Word.
- **Ở mức câu (Sequence level)**, mô hình 80k ảnh cũng đạt kết quả ấn tượng nhất với Sequence CER = 0.5330 và Sequence WER = 0.7522. Mô hình 100k ảnh lại cho kết quả tệ hơn trên cả hai chỉ số, thậm chí Sequence WER tăng lên 1.8249.
- **Dữ liệu 100k ảnh không cải thiện hiệu suất, thậm chí làm giảm chất lượng ở một số chỉ số**. Diều này có thể xuất phát từ:
  - Dữ liệu bổ sung có chất lượng thấp hoặc nhiễu,
  - Mô hình bị overfitting do số bước huấn luyện không thay đổi,
  - Phân phối của dữ liệu 100k không đồng nhất với dữ liệu kiểm thử.

Tổng kết: Mô hình huấn luyện với 80k ảnh đạt hiệu suất tốt nhất và ổn định nhất trên cả ba nhóm chỉ số: mức ký tự, mức từ và mức câu. Tập 100k ảnh không mang lại cải thiện nhiều, thậm chí còn làm tệ hơn. Nhưng nhìn chung thì cả 2 mô hình đều đã cải thiện so với mô hình cơ sở.

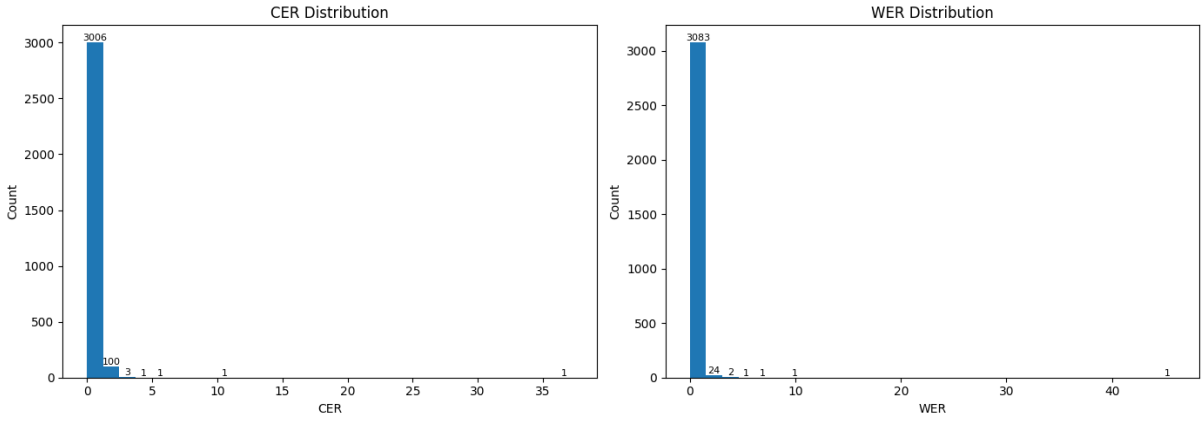
4.2 Phân phối độ lỗi của các mô hình



Hình 4: Phân phối độ lỗi của mô hình cơ sở



Hình 5: Phân phối độ lỗi của mô hình 80k



Hình 6: Phân phối độ lỗi của mô hình 100k

Nhận xét chung:

- Ở mô hình cơ sở tồn tại những mẫu có độ lỗi rất cao (tối đa có thể lên tới >120000%).
- Ở mô hình 100k thì đã cải thiện hơn so với mô hình cơ sở, nhưng vẫn tồn tại số ít các mẫu có độ lỗi cao (tối đa có thể lên tới >4000%).
- Ở mô hình 80k thì đã được cải thiện đáng kể so mới hai mô hình còn lại. độ lỗi chỉ giao động đến tối đa 400%.

5 Thảo luận

Mức độ cải thiện: Kết quả thực nghiệm cho thấy mô hình được huấn luyện với **80k ảnh** đạt mức cải thiện vượt trội so với mô hình gốc trên hầu hết các chỉ số đánh giá. Các thước đo tổng hợp như Overall CER và Overall WER đều giảm mạnh, cho thấy mô hình nhận dạng chính xác hơn ở cả mức ký tự và mức từ.

Giải thích nguyên nhân cải thiện hoặc không cải thiện Việc mô hình 80k ảnh đạt hiệu suất tối ưu nhất có thể đến từ các yếu tố sau:

- **Chất lượng dữ liệu tốt hơn**: Tập 80k có thể chứa dữ liệu đồng nhất và ít nhiễu hơn, giúp mô hình học được các đặc trưng ngôn ngữ và hình ảnh ổn định.
- **Đủ đa dạng nhưng không gây sai lệch**: Số lượng 80k ảnh có thể là mức vừa đủ để mô hình tiếp thu thông tin hữu ích mà không bị quá tải hoặc học phải nhiễu.

Ngược lại, mô hình huấn luyện với **100k ảnh** không mang lại cải thiện, thậm chí suy giảm ở nhiều chỉ số. Một số nguyên nhân tiềm năng gồm:

- **Dữ liệu mới có thể nhiễu hoặc không cùng phân phối**, khiến mô hình học phải các đặc trưng không hữu ích.
- **Overfitting cục bộ**: Mô hình có thể học lệch sang các kiểu mẫu xuất hiện ở tập 100k nhưng không phù hợp với tập kiểm thử.

Hạn chế của mô hình và quy trình huấn luyện

- **Mô hình**
  - Kiến trúc hiện tại vẫn chịu giới hạn khi xử lý các ký tự đặc biệt hoặc dấu tiếng Việt phức tạp.
  - Decoder phụ thuộc nhiều vào chất lượng embedding hình ảnh, nên dễ bị ảnh hưởng khi dữ liệu đầu vào nhiều nhiễu.
- **Quy trình huấn luyện**
  - Tối ưu siêu tham số chưa được thử nghiệm sâu, có thể chưa đạt cấu hình tối ưu.

6 Kết luận

Nhìn chung, mô hình huấn luyện với 80k ảnh đạt hiệu suất tốt nhất và mang lại sự cải thiện rõ rệt ở tất cả các mức đánh giá. Tuy nhiên, kết quả của mô hình 100k ảnh cho thấy việc mở rộng dữ liệu cần đi kèm với chiến lược huấn luyện phù hợp, kiểm soát chất lượng dữ liệu và điều chỉnh siêu tham số để đảm bảo mô hình khai thác hiệu quả lượng thông tin bổ sung.
Deepseek OCR sau khi được fine-tune thì đã phần nào đó đã cải thiện việc nhận diện các chữ cái của Tiếng Việt (không còn bị lộn sang các ngôn ngữ khác như mô hình cơ sở), tuy nhiên vẫn còn hoạt động kém ở các dấu câu, số lượng từ, ...

7 Tài liệu tham khảo

1. Haoran Wei, Yaofeng Sun, Yukun Li, DeepSeek-OCR: Contexts Optical Compression, 2025.
2. unsloth, DeepSeek-OCR: How to Run & Fine-tune, url, access time: 11/12/2025.
3. Nghia Hieu Nguyen, Duong T. D. Vo, Kiet Van Nguyen, UIT-HWDB: Using Transferring Method to Construct A Novel Benchmark for Evaluating Unconstrained Handwriting Image Recognition in Vietnamese, December 2022.
4. Khuong Xuan Dinh, deepseek_OCR finetuning, Kaggle Notebook.

8 Lời cảm ơn

Chúng em xin cảm ơn thầy *Nguyễn Hồng Bửu Long*, thầy *Dinh Điền*, và thầy *Lương An Vinh* đã giúp đỡ, và giải đáp thắc mắc của em trong quá trình học cũng như trong quá trình làm đồ án này này.

Đồ án này có sự giúp đỡ chatGPT trong việc viết Latex.