

VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY  
UNIVERSITY OF SCIENCE  
FACULTY OF INFORMATION TECHNOLOGY



INTRODUCTION TO NATURAL LANGUAGE PROCESSING

---

Project 2

# FINE-TUNING DEEPSEEK-OCR WITH VIETNAMESE DATASET

---

Lecturer: PhD. Nguyen Hong Buu Long

Ho Chi Minh City, 12/2025

# Contents

<b>1</b>	<b>Project requirements</b>	<b>2</b>
<b>2</b>	<b>Task requirements</b>	<b>2</b>
2.1	Data preparation and processing . . . . .	2
2.2	Fine-tuning the DeepSeek-OCR model . . . . .	3
2.3	Experiment design and evaluation . . . . .	3
2.4	Scientific report following research paper standards . . . . .	3
2.4.1	Background . . . . .	3
2.4.2	Methodology . . . . .	4
2.4.3	Experiments . . . . .	4
2.4.4	Results . . . . .	4
2.4.5	Discussion . . . . .	4
2.4.6	Conclusion . . . . .	4
2.4.7	Source code . . . . .	5

## 1 Project requirements

The objective of this project is to study and apply fine-tuning techniques to the DeepSeek-OCR model in order to improve the recognition quality of Vietnamese text. Specifically, students are required to complete the following tasks:

- Collect and construct a suitable Vietnamese OCR dataset, including images and corresponding ground-truth text.
- Implement the fine-tuning procedure for the DeepSeek-OCR model following the official guidelines, and clearly describe all hyperparameters, techniques, and configurations used.
- Evaluate and compare the performance of the original model and the fine-tuned model on an independent test dataset.
- Analyze in detail the improvements, error patterns, and challenges in Vietnamese text recognition after training.
- Write a scientific report presenting the full methodology, experiments, results, and evaluations following the standard structure of a research paper.

## 2 Task requirements

### 2.1 Data preparation and processing

- Students may refer to the following datasets:
  - UIT-HWDB: [UIT-HWDB](#).
  - HANDS-VNOnDB: [HANDS-VNOnDB](#).

However, students are encouraged to search for other Vietnamese datasets or collect their own Vietnamese OCR data (captured images, scanned pages, printed documents, tables, forms, handwriting, ...).

- Perform necessary preprocessing: cropping, quality enhancement, size normalization, text normalization, and fixing Vietnamese encoding issues if any.
- Split the dataset into **train/validation** for fine-tuning and **test** for final evaluation.
- Students should sample and use only a subset of the dataset to ensure sufficient computational resources for training.

## 2.2 Fine-tuning the DeepSeek-OCR model

- Students must refer to the official fine-tuning guide: [Unsloth](#).
- Students can use platforms such as Google Colab, Kaggle and so on.
- Select a fine-tuning strategy.
- Clearly document all hyperparameters used:
  - Batch size, learning rate, optimizer, number of training steps (steps/epochs),
  - Augmentation methods (if any),
- Save the model checkpoint after fine-tuning.

## 2.3 Experiment design and evaluation

- Use an independent **test** set to evaluate:
  - The original DeepSeek-OCR model.
  - The fine-tuned model.
- Use appropriate evaluation metrics:
  - Character Error Rate (CER),
  - Evaluation by data type: printed text, handwriting, tables, and forms.
- Conduct an error analysis: common error types, cases where the model improves or degrades.
- Illustrate results with input/output examples.

## 2.4 Scientific report following research paper standards

The report must include the following main sections:

### 2.4.1 Background

- Brief explanation of OCR and challenges in Vietnamese OCR.
- Overview of the DeepSeek-OCR model (architecture, input/output).

### 2.4.2 Methodology

- Detailed description of the pipeline:
  - Data collection and preprocessing.
  - Training environment setup.
  - Fine-tuning details (hyperparameters, training steps, LoRA/QLoRA if used).
- Description of inference and post-processing procedures.

### 2.4.3 Experiments

- Description of sampling strategy.
- Description of dataset splits (train/val/test).
- Experimental design comparing the original and fine-tuned models.
- List of evaluation metrics and reason for their use.

### 2.4.4 Results

- Present tables of results:
  - CER of the original model,
  - CER of the fine-tuned model,
  - Comparisons by data types if applicable.
- Provide both quantitative and qualitative analysis.
- Include illustrative example outputs.

### 2.4.5 Discussion

- Evaluate the level of improvement.
- Explain reasons for improvements or lack thereof.
- Discuss limitations of the dataset, model, or training process.

### 2.4.6 Conclusion

- Summarize the key findings.



#### 2.4.7 Source code

- Submit the training and evaluation notebook or script.
- Show OCR results and direct comparisons with the original model.