Pham Hoang Ha

Contact: (+84) 096 1190 944 | phamha.feb@gmail.com | LinkedIn | Website | GitHub

PROFILE

Aspiring Machine Learning Engineer / AI Engineer with 5+ years of industry experience in data engineering and analytics, and a recent Master's degree in Language Technology. Skilled in building end-to-end ML pipelines, with hands-on experience in deep learning, NLP, OCR, and visual-language models through academic research and applied projects.

WORK EXPERIENCE:

Uppsala University

Student Research Assistant (06/2024 - 06/2025)
Research Project: *ENABLE - Enabling climate-resilient development*

- **Main responsibility:** Implemented pipelines to convert audio, video, and scanned documents of UN Climate Change Conference statements into a plain text corpus.
- **OCR:** Built pipelines to extract text from 2000+ PDF documents (Tesseract, Google Vision API).
- **Audio processing:** Developed audio processing pipelines to transcribe 100+ conference recordings (pyannote for speaker diarization, OpenAI Whisper for transcription).
- **Machine translation:** Used Google Translate and DeepL APIs to handle non-English documents.

Teaching Assistant (09/2024 - 03/2025)

- Assisted courses: Natural Language Processing Theories, NLP Applications, Machine Learning.
- Guided 20+ students during lab sessions by reviewing Python code, clarifying theoretical and technical concepts, and troubleshooting programming issues in real time.
- Designed machine learning lab exercises, encouraging hands-on understanding of core concepts.
- Designed a handbook on using the university's High-Performance Computing (HPC) clusters.

Holistics Software - Data Analyst/Analytics Engineer (12/2018 - 06/2023)

- **Data Engineering:** Developed ETL/ELT pipelines to ingest data from various data sources into BigQuery using commercial tools, orchestrators (Prefect), dbt, and custom Python scripts.
- **Reporting & Analysis:** Built SQL-based dashboards for internal metrics, performed exploratory analysis (EDA) and investigative analysis, supporting 40+ members (Management, Sales, Engineering & Product).
- **Product feature tracking:** Design product event tracking using Snowplow.
- **Technical writing:** Contributed heavily to the company's technical blog, guides, and documents.
- **Training:** Provided training for sales teams to better communicate with technical customers.
- Managed a team of two analysts and two data engineers, and established internal data processes

Quoine - Data Analyst (07/2018 - 12/2018):

Automated operation reports and produced deep-dive analysis to improve the customer support process using Python, SQL, and R.

Tiki.vn - Data Analyst (06/2016 - 05/2018)

Automated sales & marketing reports using Airflow, and produced analysis using R and SQL.

Early Career Experience (07/2014 - 06/2016): Held various intern and trainee positions.

MASTER'S THESIS

Research Project: *Evaluating Visual-Language Models for Handwritten Text Recognition on Historical Swedish Manuscripts*
In collaboration with the Swedish National Archives (Riksarkivet)

Methodology: Fine-tuned a VLM (Florence-2) and specialized models (YOLO, TrOCR) for computer vision tasks: text region detection, instance segmentation, and handwritten text recognition (HTR). Combined these models into end-to-end OCR pipelines.

Evaluation:

- Compared fine-tuned Florence-2 against YOLO using object detection and segmentation metrics: IoU, Precision, Recall, Region Coverage, Pixel Accuracy.
- Compared fine-tuned Florence-2 against TrOCR using HTR/OCR metrics: Character Error Rate (CER), Word Error Rate (WER), and Bag-of-Words (BoW) Hits/Extras.
- Evaluated end-to-end VLM-based and classical pipelines using HTR/OCR metrics.

Results: The VLM-based pipeline outperforms the classical pipeline (CER: 11% vs. 17%; WER: 23% vs. 31% respectively).

Project: visual-language-models-htr
Demo: HTR with VLM

EDUCATION:

Uppsala University
Master of Arts in Language Technology. Graduation: 06/2025

Vietnam National University
Bachelor of Arts in Business English. Graduation: 07/2014

TECHNICAL SKILLS

- **Programming:** Python, R, SQL, Shell script, Java (basic)
- **Machine learning:** scikit-learn, PyTorch, Hugging Face, Transformers, Gradio, Streamlit
- **Database:** PostgreSQL, BigQuery, Snowflake, Redshift
- **Data visualization:** Google Data Studio, Tableau, ggplot2, matplotlib, Vega-Altair
- **Data modeling & transformation:** Holistics, dbt
- **Workflow orchestration:** Airflow, Prefect, Dagster
- **Dev tools:** git, GitHub, SSH, Terraform, Docker, Slurm, Linux/Unix, FastAPI
- **Cloud platforms:** Google Cloud Products (GCP), Amazon Web Services (AWS)

OTHER SKILLS

- English: Proficient (C1 - C2)
- Japanese: Beginner (A1)
- Swedish: Beginner (A1)