VO HOANG BAO DUY

v.hbaoduy@gmail.com +84776562199 Ho Chi Minh City, VietNam vhbaoduy

TECHNOLOGY KNOWLEDGE

Programming — Python, TypeScript, and Go.

AI/Machine Learning — PyTorch, TensorFlow, Scikit-learn, advanced ML algorithms, computer vision, signal/speech processing, natural language processing (NLP), ONNX, and TensorRT.

GenAI & LLMs — LangGraph, LangChain, RAG, GraphQL, Vector Databases, Agentic Frameworks.

Engineering & MLOps — System Design, Microservices, RESTful APIs, FastAPI, Databases (SQL, MongoDB, MySQL, Redis, BigQuery), Message Queues (RabbitMQ, MQTT) , Docker, Kubernetes, Airflow, Triton Inference Server, ClearML.

Data & Analytics Tools — Apache Spark, Polars, Pandas, NumPy, DVC.

EDUCATION

Master of Science in Computer Science 2024 – Present
VNU, Ho Chi Minh University of Science (*VNU - HCMUS*) Ho Chi Minh City, Vietnam
- GPA of **3.86/4.0.**
- **Thesis** "Speech Emotion Recognition with Multi-Scale Feature Network", paper accepted at Interspeech 2026.

Bachelor of Science in Computer Science 2019 – 2023
VNU, Ho Chi Minh University of Science (*VNU - HCMUS*) Ho Chi Minh City, Vietnam
- **Excellent degree**, GPA of **3.89/4.0**.
- **Thesis** "Text-dependent speaker verification using deep neural networks", paper presented at ACIIDS 2023.

WORK EXPERIENCES

AI Engineer Nov 2025 – Present
Qualgo Technologies Ho Chi Minh City, Vietnam
- Spearheaded **the 0-to-1 development of an internal Agent Development Kit,** architecting a ReAct agent pattern, a multi-agent reasoning framework, and LLM-based feature evaluation pipelines.
- **Engineered SLM-powered document processing workflows** within an agentic system, driving a >75% task success rate.
- Built **a multi-tier memory architecture** integrating episodic task tracking and semantic long/short-term memory to optimize context window management and achieve **>80%** retrieval recall@5.
- **Maintained and optimized production vLLM inference services**, maximizing throughput, reducing latency, and ensuring high availability for downstream agentic workflows.

Skills: TypeScript, Python, Qdrant, GraphQL, Langfuse, ONNX, OpenAI, LLM, SLM.

Data Scientist / ML Engineer Dec 2024 – Nov 2025
GiaoHangNhanh (*GHN*) - *SCommerce* Ho Chi Minh City, Vietnam
- Designed an **end-to-end ML pipeline** and **RESTful APIs** to serve forecast models, predicting warehouse volume, weight, and lead times with high accuracy (**5-15% SMAPE**) for ***Early Warning System & Operation Planning*.**
- Built an anomaly detection algorithm to proactively identify low-performance days, achieving **>85% precision** and **75% recall**.
- Optimized logistics by developing **route-clustering algorithms** (**>80% truck occupancy**) and engineering **what-if scenario analyses** for headcount and capacity planning.
- Architected an **agentic LLM system** using **LangGraph** for complex decision-making and action-taking and handling complex orchestration for **what-if analysis**.

Skills: Python, TensorFlow/PyTorch, Modeling, Spark, Polars, SQL, FastAPI, Airflow, Linux, Docker, Kubernetes, LangGraph, LangChain, LLM,

AI Engineer (Contract) Jun 2023 – Jun 2025
viAct Remote
- Engineered an **automated video inference service**, deploying YOLO and SAM models via **Triton Inference Server** (ONNX/TensorRT) and reducing deployment time by **30%** across **K8s** and edge devices.
- Built an **end-to-end streaming pipeline** for large-scale video analytics, maintaining a steady **5-15 FPS** throughput from data ingestion to final result.
- Established a **CI/CD pipeline** and maintained a core video analytics library, reducing new AI model integration time into the serving platform by **50%**.
- Designed an **ML lifecycle platform** and automated evaluation pipelines to continuously manage datasets and track AI model performance.
- Developed secondary ML models and tracking algorithms to **reduce false positive alerts** from the main inference module, improving system accuracy.

Skills: Python, ClearML, Parallel Programming, PyTorch, Numpy, FastAPI, Computer Vision, Deep Learning, MongoDB, MySQL, Redis, AWS S3, DVC, Triton Inference Server, System Design, RabbitMQ, MQTT, Linux, Docker, Kubernetes.

Algorithm Engineer Oct 2022 – May 2023
Fossil Group Ho Chi Minh City, Vietnam

- Developed and **optimized real-time algorithms** for *step count* and *human activity recognition* using wearable device accelerometers.
- Transformed raw signal data into **15+ effective features** for building ML models, achieving **>95% accuracy** and contributing to reducing the **error rate below 5%** in the step count algorithm.

Skills: Python, MATLAB, Machine Learning, Data Analysis, Feature Engineering, Signal Processing, Time Series, and Statistical Analysis

PROJECTS

Lightweight model for Speech Emotion Recognition Oct 2025 – Sep 2026

- **Architected the SETEAB framework from the ground up**, designing a novel multiscale architecture optimized for **highly efficient extraction** of temporal emotional features.
- **Achieved massive size and computational reductions**, requiring **only 0.4M to 0.5M parameters** and **0.06G to 0.12G FLOPs**. This drastically reduced computational cost compared to the wav2vec 2.0 baseline (95M parameters / 33.53G FLOPs).
- **Designed quantitative benchmark frameworks** to rigorously evaluate performance across **five diverse datasets**, achieving a peak average Unweighted Accuracy (UA) of **47.69%** and an average F1 score of **45.23%**.
- **Successfully accepted and published at Interspeech 2026**, validating the framework's novel architecture and competitive accuracy-efficiency trade-offs through rigorous peer review.

Skills: PyTorch, Signal Processing, Deep Learning, Audio Feature Extraction

Historical Document Layout Analyzer for Sino-Nôm OCR Alignment Jul 2025

- **Evaluated** deep learning object detection models to establish an automated layout analysis pipeline for complex historical texts using **PaddleOCR (PP-Structure V3)** and **YOLO11**.
- **Proved** that the Transformer-based **PP-DocLayout-L** model outperforms YOLO variants, achieving a peak performance of **0.88 AP@50** and **0.68 AP@50-90** on the SCUT-CAB dataset.
- **Built** end-to-end data preprocessing and image augmentation pipelines, integrating **Label Studio ML Backend** for automated pre-labeling, significantly reducing downstream OCR error rates.

Skills: Python, PyTorch, PaddleOCR, Label Studio, OpenCV.

AI-Powered Resume Evaluation Engine (Multi-Agent System) Sep 2025

- **Evaluated** system generalization across clean and complex, out-of-domain datasets to expose model performance boundaries under data noise.
- **Proved** traditional feature engineering (**PCA + SVR** with **Split CV–JD**) maintains superior stability on complex data, leading regression baselines with a **2.37 MAE** and **76% Acc@3**.
- **Mitigated** zero-shot LLM degradation on complex structures by deploying an "**Extraction and Evaluation**" pipeline with **Gemma 3N 2B**, securing **44% Acc@1** and **80% Acc@3**

Skills: Deep Learning, NLP, PyTorch, LangGraph, LLM.

AWARDS AND CERTIFICATIONS

Awards — Semester scholarships (2019-2023), Full scholarship (2019), and Half scholarship (2020)

Certifications — Deep Learning Specialization(Oct 2021)

PUBLICATIONS

EDWARD-NMT:Elastic Dimensions With Adaptive Routing and Dual-consistency for Neural Machine Translation Aug 2026
Conference: EMNLP 2026 Findings.
Authors: Edward Nguyen*, Thai Tran*, Kiet Anh Hoang*, Duy Hoang*, Duy Vo, Khanh Le, Vu Le.

SETEAB: Multiscale approach with Squeeze-and-Excitation Temporal Enhanced Aware Block for Speech Emotion Recognition Sep 2026
Conference: Accepted to INTERSPEECH 2026,
Authors: *Duy Vo*, *Kiet Anh Hoang*, *Hao Do*.

ViP-VL: Vietnamese Self-supervised Speech Pretraining Model with Vector-Quantization Learning Sep 2026
Conference: Accepted to INTERSPEECH 2026
Authors: Khanh Le*, Kiet Anh Hoang*, Bao Nguyen*, Duy Vo*, Dung Vo, Thai Tran, Linh Pham, Khoa D Doan.

An Effectiveness of Repeating a Spoken Digit for Speaker Verification Sep 2023
Conference: ACIIDS 2023.
Authors: Duy Vo*, Si Minh Le*, Hao Duc Do, and Son Thai Tran.