Nguyen Nhat Truong

Di An City, Binh Duong Province, Vietnam | truongnn20022003@gmail.com | +84 325762105

linkedin.com/in/nhattruongnguyen20022003/ | github.com/ChaosAIVision/

Summary

AI Engineer with a strong focus on Generative AI, including LLM training/inference, diffusion pipelines, and agent systems. Proven ability to develop and implement AI applications in image recognition and generative models. Experienced in optimizing training pipelines, model evaluation, and deployment using technologies like VLLM and Triton Inference Server. Seeking to leverage expertise in AI/ML to drive innovative solutions.

Education

FPT University Ho Chi Minh **Oct 2021 – Jun 2025**
Bachelor's Degree in Artificial Intelligence *Ho Chi Minh City, Vietnam*

- **GPA:** 7.43/10
- **Relevant Coursework:** Machine Learning, Deep Learning, Natural Language Processing, Computer Vision

Experience

AI Engineer, Pythera AI **Aug 2024 – Oct 2025**
Ho Chi Minh City, Vietnam

- **Agent System Development:**
  - Designed and implemented a **multi-agent chatbot system** using LangGraph and custom MCP tools to analyze trader behavior, track trading performance, and provide adaptive trading advice.
  - Built RAG-based workflows integrating structured data retrieval and reasoning, improving response consistency and contextual accuracy.
  - Developed workflow orchestration with Windmill to enable agent communication, parallel task execution, and cost-efficient self-hosting.
- **Dataset Engineering:**
  - Collected and processed the **Medical-O1-Reasoning** dataset for Vietnamese medical reasoning tasks.
  - Collect and preprocessed the **Stable Diffusion Inpainting 1.5** dataset for image editing tasks.
- **Model Training & Optimization:**
  - Fine-tuned DeepSeek-R1-0528-Qwen3-8B (8B) with LoRA on a 16GB GPU, improving reasoning and Q&A accuracy in Vietnamese medical contexts.
  - Trained and optimized **Stable Diffusion Inpainting 1.5** on low-end GPUs (<12GB VRAM) for object insertion, removal, and white balance correction in interior design datasets.
  - Achieved SSIM 0.833, FID 14.99 while maintaining training stability and minimal VRAM usage.
- **Deployment & Inference Optimization:**
  - Quantized DeepSeek-R1 8B model to FP8 and deployed via **VLLM**, achieving **59 TPS** inference on RTX 5090.
  - Converted Stable Diffusion Inpainting 1.5 to **TensorRT** and deployed on **Triton Inference Server**, reducing image generation latency to **6s per image** with 10GB VRAM.

AI Engineer Intern, QAI – FPT Software **Jan 2024 – Apr 2024**
Quy Nhon, Vietnam

- **Dataset Collection & Annotation:**
  - Collected and preprocessed **20k+ PPE images** from multiple industrial environments.
  - Performed bounding-box labeling and ensured annotation consistency for YOLO training.
- **Model Training & Evaluation:**

  - Trained **YOLOv8 detection model** for PPE recognition, achieving **0.95 mAP@50**.
- **Deployment Optimization:**
  - Converted trained **YOLOv8 model** to **OpenVINO**, improving inference speed by **30% on CPU**.

Projects

TrainForge **github.com/ChaosAIVision/TrainForge**
Config-driven LLM/VLM training framework with Unsloth

- YAML-based framework for **LoRA**, quantization (4/8-bit), and multi-GPU fine-tuning.
- Fine-tuned **Qwen3 4B** for **RAG reasoning** using the dataset ChaosAiVision/VI_CoT-RAG-v2.
- Achieved **0.9894 context_recall** and **0.825 faithfulness** (evaluated with **RAGAS**), running efficiently on a **minimum 8GB GPU**.

Light-Diffusion **github.com/ChaosAIVision/Light-Diffusion**
VRAM-efficient diffusion framework for object insertion

- PyTorch Lightning pipeline with Tiny VAE, 8-bit optimizer, and precomputed embeddings.
- Trains Stable Diffusion Inpainting 1.5 on a **minimum 8GB GPU** .

YOLO-AI Framework **github.com/ChaosAIVision/yolo-ai**
End-to-end real-time detection system with BentoML

- Full convert–deploy pipeline using YOLOv8, ONNX, and BentoML.
- React-based UI for image/video streaming; achieves **18–20 FPS**.

Skills

Model Training & Optimization: Experienced in training and fine-tuning models efficiently using **PyTorch Lightning**, **HuggingFace Transformers**, and **TRL**. Applied optimization techniques with **bitsandbytes**, **PEFT**, and **Unsloth** to reduce VRAM usage and accelerate training through quantization, LoRA, and gradient checkpointing.
Inference Optimization & Self-Hosting: Designed and optimized inference pipelines using **VLLM**, **TensorRT**, and **Triton Inference Server** for large-model serving. Exported models to **ONNX** for multi-platform deployment and compiled them with **TensorRT/TensorRT-LLM** on CUDA GPUs and **OpenVINO** on CPUs for lightweight inference. Applied LLM quantization to reduce VRAM usage and increase throughput, enabling stable self-hosted model serving with low latency and optimized cost.
Agent System & Workflow Development: Experienced in developing multi-agent systems using **LangGraph**, **LangChain**, and custom **MCP tools**. Built RAG-based pipelines and reasoning workflows with **Windmill** for efficient orchestration, context tracking, and reduced redundant computation in production environments.
Databases: PostgreSQL, SQLite
Tools & Environment: Git, Docker, Cursor, Claude Code, Linux
Languages: Vietnamese (Native), English (Good proficiency)