Quan Nguyen

quanhnguyen232@gmail.com College Park, MD QuanHNguyen232 QuanHNguyen232

Education

University of Maryland, College Park, Computer Science — College Park, MD
Jan 2024 – Jan 2026

Gettysburg College, Computer Science — Gettysburg, MD
Jan 2020 – Jan 2024

- Phi Beta Kappa Society member
- David Wills Scholarship recipient

Experience

Venera AI, Machine Learning Engineer Intern - LLM Post-training & ML System — New York, NY
Feb 2025 – Jan 2026
1 year

Developed and deployed a scalable and efficient LLM post-training and ML system to improve the accuracy and efficiency of the model.

- Fine-tuned (SFT and distillation) Qwen3 to compress knowledge using QLoRA + DeepSpeed ZeRO-3
- Engineered a high-throughput inference for LLM; support prefix/KV caching, continuous batching.
- Deployed LLM on TPU v5 (2x4 pod, 8 chips), establishing a cost-efficient alternative to GPUs.
- Rebuilt data pipeline Spark with Ray Data, managed via Airflow, 3x throughput to 1M+ tok/min.
- Consolidated CI/CD (GitHub Actions + Terraform + Helm) and monitoring (Grafana + Prometheus).

Adobe, Machine Learning Engineer Intern - AI Agent — San Jose, CA
May 2025 – Aug 2025
4 months

Developed and deployed a scalable and efficient AI agent to improve the accuracy and efficiency of the model.

- Developed Voice Agent features combining planning, speech recognition, and emotion-aware TTS.
- Built AI-agent using LangChain, LangGraph, and MCP to integrate into Adobe multi-agent system.
- Deployed production models via vLLM, Ray Serve, FastAPI, integrated with Kubernetes and ArgoCD.

VCCorp Corporation, Machine Learning Engineer Intern - Recommendation System — HCMC, Vietnam
June 2024 – Aug 2024
3 months

Developed and deployed a scalable and efficient recommendation system to improve the accuracy and efficiency of the model.

- Developed a scalable and efficient recommendation system to improve the accuracy and efficiency of the model.
- Developed a scalable and efficient recommendation system to improve the accuracy and efficiency of the model.

Awards

ICPC Participant — Jan 2022
ICPC is a competitive programming contest for university students.
ICPC
icpc.io

Publications

Predicting Perceived Music Emotions with Respect to Instrument Combinations
Music Emotion Recognition has attracted a lot of academic research work in recent years because it has a wide range of applications, including song recommendation and music visualization. As music is a way for humans to express emotion, there is a need for a machine to automatically infer the perceived emotion of pieces of music. In this paper, we compare the accuracy difference between music emotion recognition models given music pieces as a whole versus music pieces separated by instruments. To compare the models' emotion predictions, which are distributions over valence and arousal values, we provide a metric that compares two distribution curves. Using this metric, we provide empirical evidence that training Random Forest and Convolution Recurrent Neural Network with mixed instrumental music data conveys a better understanding of emotion than training the same models with music that are separated into each instrumental source.
Nguyen, Viet Dung, Nguyen, Quan H., Freedman, Richard G.
ojs.aaai.org/index.php/AAAI/article/view/26910

Skills

Machine Learning

Languages

Vietnamese
Native speaker

English
Fluent

Interests

Machine Learning

Certificates

Fundamentals of MCP July 2025

Projects

High-Performance Distributed Training (HPC) Jan 2025 – Jan 2025
Accelerated 3D scene reconstruction training by 50% by implementing distributed training and profiling using C/C++, CUDA, PyTorch DDP with MPI protocol across multi-GPU HPC clusters.

- Implemented distributed training and profiling using C/C++, CUDA, PyTorch DDP with MPI protocol across multi-GPU HPC clusters.
- Optimized the training process by 50% by implementing distributed training and profiling using C/C++, CUDA, PyTorch DDP with MPI protocol across multi-GPU HPC clusters.