Nguyen Van Phuoc

Ho Chi Minh City (+84)376.781.543
vectornguyen76@gmail.com linkedin.com/in/vectornguyen76 github.com/vectornguyen76

Education

University of Technology and Education **Sep 2018 – Sep 2022**
Mechatronics Engineer *Ho Chi Minh*

Experience

AI Engineer | *AI Center - FPT Software* **Jun 2022 - Present**

- Developed cutting-edge RAG chatbots capable of processing diverse document types including images, tables, and text, significantly enhancing information retrieval and user interaction across multiple domains.
- Implemented multi-agent systems to facilitate dynamic conversations between humans and bots, effectively extracting information and routing intents in IT support and customer service scenarios.
- Engineered multilingual multi-agent solutions for Japanese and German markets, adeptly handling cultural nuances and achieving high performance, resulting in positive client feedback and expanded use cases.
- Reduced new project setup time by 35% through the creation of backend, CICD pipeline, IaC and AI service templates, streamlining demo and production implementation.
- Ensured scalability, performance, and security by deploying microservices using Docker, Serverless and Kubernetes.
- Proactively monitor, analyze, troubleshoot, and optimize the performance of AI models in production.

Back End Engineer | *Hao Phuong Corp* **Jan 2022 - Jun 2022**

- Utilized DevExpress and ASP.NET Core Web API framework to create a real-time smart Farm management solution, enabling efficient data collection, control, and visualization for informed decision-making.
- Collaborated effectively with mobile teams to design a comprehensive ERD database model, ensuring data integrity and consistency across platforms.

Side Projects

Search Engine System | *github.com/vectornguyen76/search-engine-system* **Dec 2023**

- Architected and developed a scalable microservices system handling data from 100,000 Shopee products, demonstrating expertise in large-scale data processing and system design.
- Implemented an Image Search Engine using asynchronous programming, gRPC, Vector Database, TensorRT, and Dynamic Batching, achieving 200 RPS with 100ms p95 latency.
- Engineered Full-text Search and Autocomplete functionality using Elastic Search with custom scoring scripts, improving search relevancy by 35%.
- Established continuous integration and deployment (CI/CD) using Docker, GitHub Actions, Ansible and CloudFormation. Utilized Helm Charts to deploy the project on EKS in AWS.

Face Recognition System | *github.com/vectornguyen76/face-recognition* **Sep 2023**

- Developed a real-time face recognition system using One-Shot Learning, enabling accurate identification and authentication for multiple users simultaneously. This approach significantly reduced the need for extensive training data while maintaining high recognition accuracy.
- Engineered a scalable face detection, tracking, and recognition pipeline optimized for high performance. Utilized Redis Streams for efficient data streaming, Vector Database for rapid similarity search, and Triton Server for serving deep learning models. Achieved scalability with the system supporting 200 requests per second (RPS) for face detection and 100 RPS for face recognition.

Resume Ranking | *github.com/vectornguyen76/resume-ranking* **Apr 2023**

- Innovated a resume ranking system leveraging the advanced language processing capabilities of large language models.
- Utilized prompt engineering to extract key insights from resumes and create accurate rankings.

Technical Skills

Languages: Python, Javascript, C#, C, SQL
Machine Learning/Deep Learning: LLMs, Natural Language Processing, Computer Vision
Cloud DevOps: AWS, Docker, Kubernetes, Serverless, CircleCI, Github, Github Actions, Ansible, Linux
Frameworks/Libraries: PyTorch, Tensorflow, Numpy, FastAPI, Flask, Triton Inference Server, Vector Database

Honors and awards

Most Valuable Player 2022 **Jan 2023**
Second Prize in Quy Nhon AI Hackathon 2022 **Sep 2022**
First Prize in Defend Graduation Project **Jun 2022**
Top 4 in Self-Driving Car Competition **May 2022**