Hoang Pham

linkedin.com/in/hoangph3

Email : phamhoanghxh1@gmail.com

Mobile : +84 339 362 666

I like building large-scale ML pipeline!

Experience

- **Machine Learning Engineer (Mid-level)** Jul 2020 - Present
  Viettel Cyber Security *Hanoi, Vietnam*
  - **E2E Pipeline Design**: Architected and implemented scalable machine learning pipelines to automate the data ingestion, processing, model training, evaluation, and deployment processes.
  - **Model Training and Tuning**: Implemented automated model training workflows using frameworks like TensorFlow, Torch, or Sklearn. Leveraged hyperparameter tuning strategies to optimize model performance.
  - **Monitoring and Maintenance**: Established monitoring solutions for model performance and data drift, ensuring the reliability and accuracy of models. Implemented retraining mechanisms based on new data or model decay.
- **Software Engineer AI** Jun 2022 - Present
  Freelance *Remote*
  - **Application of Pre-trained Models**: Specialized in applying pre-trained models for tasks such as LLM, NLP, computer vision, utilizing frameworks like ONNX, PyTorch, and Hugging Face.
  - **Solution Architecture**: Developed robust architectures for AI-powered applications, focusing on optimizing performance and ensuring seamless integration with existing software systems.
  - **System Integration**: Successfully integrated AI functionalities into client applications on edge platforms (Jetson), leveraging technologies like ONNX and TensorRT for deployment in resource-constrained environments.

Education

- **Hanoi University of Science and Technology** Hanoi, Vietnam
  Bachelor, Automation Engineer Technology; Grade: 3.37/4.0 *Sep 2017 – Jun 2021*

Certifications

- **Deeplearning.AI**: Machine Learning Engineering for Production (MLOps) Specialization
- **Deeplearning.AI**: Deep Learning Specialization
- **Stanford Online**: Machine Learning

Programming Skills

- **Languages**: Python, NodeJS, Javascript, SQL, Spark
- **Technologies**: Jetson, ONNX, Kafka, Airflow, Kubeflow, MLFlow, Minio, K8S, PyTorch, Tensorflow

Projects

- **Security Chatbots** Mar 2024 - Present
  Viettel Cyber Security
  - **SOAR Alert Identification**: Develope chatbots to identify and categorize SOAR (Security Orchestration, Automation, and Response) alerts, enabling faster and more accurate threat detection.
  - **Automated Alert Operations**: Implement automated workflows within the chatbots to handle alerts based on predefined checklists. Utilized AI prompting to planning, execute tasks efficiently and consistently.
  - **Automated Enrichments**: Integrate chatbot with some tools to update new information continually to make decisions, avoid outdated information.
  - **Alert Reporting**: Design the chatbots to automatically summarize alert details and generate comprehensive reports, streamlining the process for end-users to review and respond to security incidents.
  - **Multi model RAG**: Extract images, text, and tables from documents, summarize, and store raw text and images along with their summaries for retrieval, then synthesize final answer from join review of images and texts.
  - **Finetuning LLMSec**: Domain-specific LLM trained on cybersecurity data (CEH tests, Atomic Red Team tests, MITRE ATT&CK documentation.
  - **Tech Stack**: HuggingFace, Airflow, Celery, Docker, SQL, Metabase, Redis, Python, Dify, Langchain, VLLM, LLama.cpp.

- **Manufacturing Chatbots** Oct 2023 – Mar 2024
  DENSO Vietnam
  - **Production Line Information Retrieval**: Develope chatbots to provide detailed production line information, such as equipment downtime and causes, by querying SQL Server databases.
  - **LLM Localization**: Hosting LLMs on-premise (llama-7b, mistral-7b)
  - **Entity Extraction with Hybrid-RAG**: Implement Retrieval-Augmented Generation (RAG) to extract named entities from user queries, improving the precision of information retrieval.
  - **Keyword Matching with Embeddings**: Using word embeddings for keyword matching, ensuring that the chatbot could accurately interpret and respond to user queries based on semantic similarity.
  - **Data Management**: Store historical chat data and interaction logs using MongoDB, enabling easy access to past queries and responses.
  - **Telegram Integration**: Integrate the chatbot with Telegram, adding features like feedback buttons, daily reports, and command-based job execution to enhance user interaction and efficiency.
  - **Tech Stack**: HuggingFace, Telegram Bot, WebHook, FastAPI, Milvus, Celery, NoSQL, SQL, Python, Kubernetes, Ollama, VLLM.

- **Alert Response Automation** Jan 2023 - Present
  Viettel Cyber Security
  - **Embedding-Based Similarity Search**: Using embedding techniques to perform similarity searches on alerts, enabling the automatic identification and closure of false positives, improving overall efficiency and accuracy in alert management.
  - **False Positive Reduction**: Implement automation for identifying and managing false positives, significantly reducing the manual workload for security teams.
  - **Incident Response Automation**: Develope systems to automate incident response processes, leveraging historical operating data to streamline investigations and reduce response times.
  - **Tech Stack**: Faiss, FastText, SQL, NoSQL, Flask, Kafka, Python, Nomad, Consul.

- **People Counting** Sep 2023 - Jan 2024
  Indochina Plaza Hanoi
  - **Multi-Camera Tracking**: Develope a system for tracking individuals using both single and multi-camera setups, with integration via RTSP streams from cameras deployed in shopping malls.
  - **Event-Driven Architecture**: Implement an event-driven system using Redis and Kafka for message queuing, enabling real-time processing and scalability across distributed devices.
  - **Person Re-Identification (ReID)**: Applying embedding methods for person re-identification, utilizing similarity search and re-ranking techniques to accurately track and count individuals across different camera feeds.
  - **Distributed Edge Processing**: Deploying the system on multiple NVIDIA JetsonNX devices, optimizing the pipeline for real-time performance on edge hardware.
  - **Tech Stack**: NVIDIA Jetson, Pytorch, ONNX, Kafka, MongoDB, Clustering, Python, Docker.

- **Face KYC** Sep 2022 - Feb 2023
  Telehouse Vietnam
  - **Camera Management**: Develope a system for capturing images from tablets or IP cameras, ensuring high-quality input for facial recognition processes.
  - **Model Serving**: Hosting ONNX-based face detection and recognition models on NVIDIA Jetson Xavier devices, leveraging CUDA for accelerated performance.
  - **Door Control System**: Implement a door control mechanism using ESP32 and GPIO with UART protocol, allowing for secure access control through USB port communication.
  - **Licensing Management**: Design a licensing system using JWT with expiration time to manage and secure API access, ensuring compliance with security protocols.
  - **Backend Development**: Manage backend processes including user management, CR handling, and integration with the CASDM Broadcom system, providing a robust and scalable infrastructure for the solution.
  - **Tech Stack**: NVIDIA Jetson, ESP32, Facial Recognition, Flask, NodeJS, Python, Docker.

- **Typing biometrics** Jan 2022 - Mar 2022
  Viettel Cyber Security
  - **Typing Behavior Authentication**: Develope a system to authenticate users based on their typing patterns during login and continuous work sessions, enhancing security beyond traditional methods.
  - **Metric Learning Implementation**: Applying metric learning techniques to create an encoder that transforms typing behavior into distinct user features, enabling accurate identification.
  - **User Profile Management**: Implement a system to save typing feature embeddings into user profiles during registration, with OTP verification for added security.
  - **Behavior Comparison**: Design algorithms to compare new typing behaviors with stored user features, ensuring consistent and reliable authentication.
  - **Tech Stack**: Tensorflow, Flask, SQL, Clustering, Python.

- **Transaction Fraud Detection** Sep 2021 - Dec 2021
  Viettel Cyber Security
  - **Pattern Extraction and Time Series Conversion**: Extract patterns from abnormal transaction behaviors and converted them into multiple time series for detailed analysis.
  - **Real-Time Behavior Comparison**: Implement algorithms to compare real-time transaction behaviors with learned abnormal patterns, using techniques such as dynamic time warping and longest common subsequence for accurate detection.
  - **Behavior Comparison**: Applying methods like Discrete Fourier Transform (DFT) and Piecewise Aggregate Approximation (PAA) to approximate time series data, enhancing the efficiency of pattern recognition and fraud detection.
  - **Tech Stack**: Statsmodels, Matching Algorithms, Kafka, Pandas, Numpy, Python.