

BÙI HOÀNG ANH

Thực tập sinh AI Engineer

📅 21/10/2004
📞 0783128994
📍 Vũ Tông Phan, Thanh Xuân, Hà Nội.
✉️ hoanganhbui2110@gmail.com
🌐 https://github.com/PearlHanh

MỤC TIÊU NGHỀ NGHIỆP

Mong muốn được thực tập trong lĩnh vực khoa học dữ liệu và trí tuệ nhân tạo để vận dụng kiến thức đã học được vào giải quyết các vấn đề thực tế, nâng cao khả năng khai thác, đồng thời tăng khả năng phân tích, xử lý dữ liệu bằng các công cụ như Python, SQL và các thư viện học máy. Đồng thời học hỏi từ các chuyên gia đầu ngành, được làm việc trong môi trường chuyên nghiệp và đóng góp giá trị cho các dự án thực tế và phát triển sự nghiệp lâu dài trong ngành.

HỌC VẤN

Chuyên ngành: Khoa học máy tính - IT1

Đại học Bách Khoa Hà Nội

2022 - NAY

Học phần chuyên sâu: Khoa học dữ liệu và Trí tuệ nhân tạo

DỰ ÁN CÁ NHÂN

Hệ thống thu thập, xử lý, lưu trữ, phân tích dữ liệu biến động giá của sàn chứng khoán

10/2025 - NAY

Mô tả:

- Dự án tập trung nghiên cứu, xây dựng pipeline tự động để thu thập dữ liệu từ RESTAPI (Khoảng 60 triệu bản ghi gồm thông tin các phiên giao dịch của các loại tiền mã hóa) trên data.binance.vision và Coinbase, biến đổi theo cấu trúc chung và biểu diễn dữ liệu phục vụ phân tích.
- Dự án sử dụng Airflow triển khai trên server Compute Engine để tự động thu thập, lưu trữ và biến đổi dữ liệu. Sau khi gọi API dữ liệu sẽ được đưa vào dịch vụ lưu trữ Google Storage Service, sau đó được làm sạch và biến đổi bằng Spark và được đưa lên dịch vụ data warehouse là Big Query. Dữ liệu được trực quan hóa qua Streamlit.
- Dự án đồng thời xử lý streaming bằng Kafka với nguồn dữ liệu từ WebSocket của Binance.

Công nghệ sử dụng:

- Kubernetes:** Triển khai và điều phối các microservices, đảm bảo auto-scaling và fault tolerance
- Compute Engine:** dịch vụ máy chủ ảo trên đám mây.
- Apache Airflow:** lập lịch và quản lý luồng.
- Google Storage:** data lake chứa dữ liệu thô.
- Google BigQuery:** data warehouse chứa dữ liệu đã được làm sạch và biến đổi.
- Spark:** làm sạch, biến đổi và chuẩn bị dữ liệu cho mục đích phân tích.
- Kafka:** streaming data.
- Streamlit:** hỗ trợ trực quan dữ liệu

Mô tả:

- Áp dụng các kỹ thuật học máy để phân loại cảm xúc (7 nhãn) từ ảnh khuôn mặt sử dụng dữ liệu từ Kaggle (RAF-DB).
- Tiền xử lý dữ liệu bao gồm chuẩn hóa kích thước, chuyển sang grayscale, trích xuất đặc trưng bằng HOG, xử lý mất cân bằng bằng kỹ thuật tăng cường dữ liệu có chọn lọc.
- So sánh hiệu năng của các thuật toán như KNN, Decision Tree, Random forest, Logistic Regression, SVM và CNN.
- Tối ưu mô hình với Grid Search và đánh giá qua Accuracy, F1 Score và ma trận nhầm lẫn.

Công nghệ sử dụng: Python (Scikit-learn, NumPy, Matplotlib, cv2), Tensorflow (Keras).

Kết quả: Mô hình SVM đạt độ chính xác cao nhất trên cả tập Train (85%) và tập Test (75%), có khả năng tổng quát hóa tốt hơn so với các mô hình khác.

DỰ ÁN NHÓM

Real-time-traffic-sign-recognition

Vai trò: thu thập, tiền xử lý dữ liệu, fine-tune model yolo và tối ưu hiệu suất

9/2025 - 10/2025

Mô tả:

- Xây dựng hệ thống **nhận diện biển báo giao thông thời gian thực** trên camera hành trình.
- Thu thập dữ liệu, chuẩn hóa theo định dạng của YOLO.
- Fine-tune YOLOv12** trên tập biển báo đã tiền xử lý; kết hợp **vision-language model (VLM)** để hỗ trợ phân loại ý nghĩa của biển báo.
- Chuyển mô hình sang **TensorRT**, giảm latency suy luận
- Tối ưu pipeline OpenCV + CUDA và loại bỏ bước xử lý dư thừa giúp hệ thống có khả năng chạy real-time (≥ 25 FPS) trên GPU/Edge device.
- Công nghệ sử dụng:** YOLOv12, Ultralytics, Albumentations, Supervision, PyTorch, NumPy, OpenCV.

Kết quả: Nhận diện tốt các biển báo giao thông với mAP50-95 đạt 87.3 điểm, có khả năng nhận diện tốt, hoạt động real-time ổn định.

KỸ NĂNG

Kỹ năng chuyên ngành

- Ngôn ngữ lập trình:** Python
- Framework:** NumPy, Matplotlib, Pandas, Skikit-learn, Tensorflow, PyTorch, CV.
- Database:** MySQL, SQL Server, NoSQL
- Tools:** Github, LaTex, streamlit, k8s.

Kỹ năng mềm

- Giao tiếp:** Trôi chảy, tốt
- Làm việc nhóm:** tích cực đóng góp
- Tư duy:** Logic, giải quyết vấn đề
- Tiếng Anh:** Mức khá