NGUYEN THAI BAO – Data Engineer

Phone: +84 898 489 885 Email: baonguyenthai580@gmail.com Github: https://github.com/baobao1911

EDUCATION

University of Science, Vietnam National University Ho Chi Minh City

Bachelor of Science in Computer Science, High-Quality Program October 2020 – Present (Expected graduation date)

GPA: 2.94

- Thesis : “Identyfy Prostate Cancer from Pathology Images using Pyramid Scene Parsing Network”
  Developing upon the PSPNet model, my thesis focused on refining segmentation accuracy for prostate cancer using WHO slide images from the MCCAI conference and real data from the University of Medicine and Pharmacy at HCMC. Achieving superior performance, exceeding the original model's mIoU and Dice metric, underscores my capacity for innovative improvement within machine learning and medical imaging.
  (Scored 9.3/10)
- Achivement:
  - Engaged in scientific inquiry involving medical datasets.
  - Ranked in the Top 20 of “DAZONE - Cuộc thi Phân Tích Dữ Liệu 2023”.

SKILLS

Technical: Python, PostgreSQL, Pytorch, MLflow, Apache Spark, Apache Hadoop, Apache Airflow, Apache Kafka, Apache Flink, Selenium, BeautifulSoup, Power BI .

Skills: Machine Learning, Data Processing, Data Visualization, Data Warehouse, Data Lake, Data Analysis, ETL Processes, AWS (Familiar with basic concepts).

Language: English

- Listening: Intermediate
- Writing: Intermediate
- Reading: Good
- Speaking: Basic

PROJECTS

Prostate Cancer Segmentation from pathology images

July 2023 – December 2023

- Programming language: Python
- Frameworks: Pytorch, Albumentations, Jupyter Notebook, Numpy, Math, FlashAPI.
- Github: baobao1911/Prostate-Cancer-Segmentation-from-pathology-images
- Description : This project endeavors to develop an advanced AI model for the automated identification of malignant regions within prostate cancer pathology images sourced from the prestigious MICCAI Gleason Grading Challenge. The model undertakes comprehensive enhancements to the PSPNet architecture, including meticulous preprocessing of H&E images, meticulous data balancing procedures, precise fine-tuning of the pre-existing ResNet backbone, and systematic modular improvements. The overarching objective is to achieve a notable enhancement in segmentation accuracy, thereby facilitating more precise diagnostic insights and refined treatment planning protocols.

- Tasks: Preprocessing H&E image data, data balancing, fine-tuning pre-trained ResNet backbone, implementing modular improvements based on PSPNet, training and testing, deploying model with FlaskAPI.

Features storage

February 2024 – March 2024

- Programming language: Python
- Frameworks: Apache Airflow, PostgreSQL, Apache kafka, Apache Flink.
- Github: baobao1911/Features_storage
- Description: This project is focused on the conception and implementation of an advanced feature store system meticulously tailored for the management and provisioning of trip records sourced from the Yellow Taxi fleets of New York City. The system is architected to accommodate the distinctive requirements inherent in each data stream, thereby ensuring unparalleled performance and operational efficiency. Through the adept utilization of leading-edge technologies and platforms including PySpark, PostgreSQL, Flink, Kafka, DBT, and Airflow, the endeavor is poised to deliver a resilient and dependable solution to meet the exigencies of the domain.

Web scraping and Visualize

February 2023 – April 2023

- Programming language: Python
- Frameworks: Selenium, Pandas, Matplotlib, Jupyter Notebook.
- Github: baobao1911/Web-Scraping-and-Analysts
- Team size: 3
- Description: This project entails the utilization of the Selenium library to systematically extract data from a designated web page (WhoScored.com). Subsequently, the acquired data is meticulously processed to conform to an appropriate format conducive to visualization, employing the Matplotlib library. The outcome comprises a comprehensive graph accompanied by insightful annotations elucidating pertinent aspects of the extracted data.
- Tasks: Data crawling from website utilizing Selenium and analysis of two specific aspects.

CERTIFICATIONS

- Data Visualization University of Illinois at Urbana-Champaign April 2023
  coursera.org/share
- Text Retrieval and Search Engines University of Illinois at UrbanaChampaign May 2023
  coursera.org/share