John Pham

Data Scientist | *Machine Learning Engineer*

johnphamshine@gmail.com +1 949 328 4957

Summary of experience:

- Strong experience in Python scripting and build cloud/desktop applications and its libraries for ML models/servers (PyTorch, PySpark, Pandas, Flask/Django, asyncio)
- Strong understanding and experience in various ML models including Computer Vision and NLP recognition/generation ones – ML/DL models back behind by from-scratch training and fine-tuning different models with correct data engineering and science
- Strong experience in content generation models, and services and combining them in one pipeline with good prompt engineering and parameter adjustment – transformers, stable diffusion models, GANs, Midjourney, PlayHT, OpenAI API, Llama, etc
- Strong experience in deployment of built ML models onto cloud servers and edge devices by optimization/quantization into light-weight models and performance analysis in high insight metrics, various experimental practices comparing different state-of-the-art models' performance
- Strong experience in MLOps tools and platforms – Terraform, Nvidia Triton Server, Hugging Face, Weights and Biases
- Strong experience in using and configuring GCP services for ML training and serving instances (Cloud Function, Cloud Tasks, Monitor, Compute Engine, App Engine, Google Storage Bucket, FireStore, Redis, VPC, IAM, Security/Firewall) and GCP CLI
- Good experience in using and configuring AWS services (ECS, EC2, IAM, S3, VPC, Batch, Lambda) and AWS CLI
- Good experience in architect from scratch and a whole MLOps: resources planning and work coordination, requirements analysis and project estimation, reviewing the work commits, facing customers, team progress tracking and mentoring, creating competency requirements
- Strong experience in using and configuring the Docker platform for CI, CD process and server containerization with Kuberflow/Kubernetes
- Experience in configuring and using DBs – SQL and NoSQL DBs
- Good experience in compiling and building C++/Python applications on different platforms – Windows, Ubuntu, MacOS, Android
- Experience in different edge devices with their configuration and running instances – Raspberry-pi, different ARM chips, Stream Cameras, IR Cameras, Lidar, and Nvidia Jetson Series

Technology/Methodology:

Machine Learning, Natural Language Processing (NLP), Data Science, Data Engineering, Computer Vision, Deep Learning, OpenCV, Convolutional Neural Networks (CNN), Artificial Intelligence (AI), PySparkRL, NLP, NLU, LLMs, KNN, k-means, agglomerative clustering, linear classifiers, SVM/kernel methods, LDA, TFIDF, BERT, GPT3/3.5/4, Fasttext, Lngchain, CNN(residual-1dCNN), RNN(nmt, bidaf, seq2seq), ontology, finite automata(task completion), python, JavaScript, C++, SQL, PySpark, pandas, Tabular, scala, Kafka, polars, nltk, Gensim, spacy, Fasttext, scikit-learn, Scipy, word2vec, glove, Tensorflow, Keras, Pytorch, Statistics, Yolo, maskRCNN, GANs, OpenCV, ultrasound sensors, IR sensors, Google Cloud Platform tools, docker, Kubernetes, AWS

Work experience:

Cloudera
May 2019 – May 2024

Senior Data Scientist

- LLM-based domain specified and business information analysis assist Chat/QA bot/reporter building. Based on the agricultural and biological knowledge data and the plants' real-time status, an analysis and interaction features were built by LLM. Also, by the employee/client feedback and review for companies - business analysis reports, and question-answer assistant implementation using LLM models and Agent system. Mainly working on vector database configuration and LLM model building and its pre/post-processing.
- Played with Computer Vision models for anomaly detection and image segmentation. Implemented multi-modality VLM(Vision Language Model) on domain data.
- As a separate project, I worked on the AI server integration of a storybook content generation for children. Combined different content generation models and APIs of stable diffusion XL/midjourney wrapper, GPT4, and PlayHT for Scene Generation + Caption + Audio Generation.

  Responsibilities:
  - Statistical sentiment analysis
  - LLM pipeline design for AI analyst and business recommendation bot using OpenAI
  - Text-based regularly updating data stream chunking and vector embedding implementation
  - LLM prompt engineering support and research to enhance the response quality
  - Digging into modern RAG systems and integration experiments
  - GCP cloud function, tasks, compute engines, firestore RTDB, and buckets for ML model instances and regular updates
  - Built AI server prototype of baby storybook/audio book auto-generation by content generation models
  - YOLO-based real-time object detection and segmentation
  - Training/Inference of multi-modality Vision Language models to understand the domain images.

  Environment and Tools: GCP, Docker, Terraform, Pinecone, Supabase, Tabula, SQL, Redis, Langchain, OpenAI API, Llama, Hugging Face, A40 GPU, Cuda, PyTorch, OpenAPI, Flask, FastAPI, Streamlit, Github workflow, Jira, Mirror, midjourney, stable diffusion, ChatGPT4, PlayHT

InData Labs
April 2018 – May 2019

Senior Analyst, Senior Machine Learning Engineer

- Played with text classification models for contextual text blobs. Built and implemented email classification and data extraction systems(intents and entities).

  Responsibilities:
  - Video caption classification and clustering algorithms and models training
  - Designed document(pdf) structured data extraction system and models.
  - Configured and deployed NLP pipelines and models, quality control system (to be sure every fix improves business metrics, such as accuracy and others).
  - Optimized production-ready fast and efficient models on a distributed architecture.
  - Docker, Google Cloud Run, Function, and Tasks implementation
  - Youtube, tiktok video channels import in high frequency and large volume data import environment

  Environment and Tools: AWS, GCP, Lambda GPU Cloud, Docker, Novu, Courier, Google BERT, firebase, SQL, NLTK, Apache Airflow, PySpark, Pandas, Bitbucket, Jira

Upwork
January 2017 – March 2022

Python Engineer/ ML Engineer / Data Engineer on Several small/mid sized projects

- Worked on several projects with professional experience in Data Science and ML|DL solutions development

Projects:

- OCR model building and deployment for official identity documents. Image-to-image Conv-Deconvmodel and Tesseract were used.
- Video compression based on optical flow and pixel quantization.
- Automatic target shooting score judgment system based on object Detection and image matching algorithms.
- Recommendation system and Reinforcement model for user experience enhancement in online games.
- Betting result prediction model development for Hong Kong Horse Racing.
- Bat call detection implementation on Raspberry Pi device from sensored real-time ultrasound data.

Environment and Tools: Nvidia Jetson, Raspberry Pi, ARM, NLPU/TPU, Nvidia Triton, Cuda C++/Python, R, Matlab, Docker, Kubernetes, Apach Kafka, Stoplight, Heroku, Gitlab CI, Wix, GitHub, GCP, AWS, aiohttps, OpenCV, media-pipe, python asyncio, onnx, tflight, Roboflow, Bitbucket, Jira, Microsoft Azure, Linux, MacOS, Android, Blender

Machine Learning Engineer at FF-Group project

- Machine Learning Deep Learning Engineering for real-time vehicle detection and license plate recognition

  Responsibilities:
  - Real-time object detection by CNN models - yolo and SSD
  - Trained model compilation and optimization on device architectures - arms, tpus and dlpus
  - Model inference latency control on Axis camera sensor chips
  - Data science and data engineering for LPD, python backend engineering for the data management and serving by PySpark and CKAN
  - Linux compilation and docker operation

  Environment and Tools: CKAN, PySpark, Falsk, Axis Cameras, Arm chips, Docker/Docker Hub, Larod Inference Server, CMake, C/C++, OpenCV, Linux, Bitbucket

WarmElsa
January 2018 – February 2019

Tech Co-Founder

- From scratch building an AI vision and time-series model in the startup environment

  Responsibilities:
  - Lead and management of a cross-functional tech team of web, mobile, and ML members
  - Computer vision algorithms and CNN implementation for face, hands, and body landmark extraction- Gesture and action recognition for body language representation by time-series models on landmark frames
  - Integrated an avatar generation model
  - Low latency model optimization and deployment on mobile devices

  Environment and Tools: Lambda GPU cloud, media-pipe, OpenCV, ffmpeg, android, tflight, NLTK, ayncio, GitHub, ngrok, ngnix, Jira, Docker

L2 Inc.
August 2017 – May 2018

Team Lead

- Backend service development and internal data tools development

  Responsibilities:
  - Python-based web applications and desktop applications
  - Asynchronous applications and cryptography algorithms implementation for data security
  - Internal annotators, statistical data analysis and engineering tools

  Environment and Tools: pycryptodome, AES, matlab, PyQt, Django, Flask, Firebase, Google Storage Bucket, Google Cloud Compute Engine, VPN, Linux, MacOS

Yale School of Medicine *August 2013 – May 2014*	***Research Assistant, Python Engineer***

Education:

Georgia Institue of Technology *2019 – 2021*	Master's degree in Computer Science and Artificial Intelligence
Yale University *2010 – 2014*	Bachelor’s degree