Luong-Ha Nguyen

+1 (438) 928-2616 luongha.nguyen@gmail.com Montreal, QC, CANADA
hazone.me github.com/lhnguyen102 linkedin.com/in/luong-ha-nguyen-697941b4

Experienced AI/ML Engineer with 3 years in industry and 4 years in applied research. Specializes in developing AI/ML pipelines, from data preprocessing and model development to cloud deployment. Skilled in C++/CUDA, Python, and deep neural networks' uncertainty modeling. Combines engineering skills with research insights to deliver robust solutions.

SKILLS

Language	C/C++, CUDA, Python, JavaScript, React.
Machine Learning	PyTorch, TensorFlow, Numpy, Pandas, Scikit-learn, Probability & Statistics, Reinforcement Learning, Statistical Modeling, Computer Vision, Tabular Data Forecasting, Data Analysis, Data Visualization.
Developer Tools:	GDB, CUDA-GDB, Nsight Compute, Terraform, AWS, Azure, Sagemaker, Microservices, gRPC, REST API, Kubernetes, PostgreSQL, Docker, GitHub, Helm Charts, Hugging Face.

PROFESSIONAL EXPERIENCE

Machine Learning Engineer at AI Redefined | Montreal, QC, CA July, 2022 - present

- Build end-to-end machine learning pipelines for time series forecasting and anomaly detection, using reinforcement learning with human feedback to continually improve accuracy through insights from operators and asset managers, directly contributed to a 50% increase in Annual Recurring Revenue (ARR).
- Lead the development of vision-based detection algorithms for fault detection, directly contributed to a 10% increase in ARR.
- Finetune large language model (LLM) for internal chatbot including data collection, model selection, finetuning strategy, evaluation, and deployment.
- Create cloud-based tools for automating model training, hyperparameter tuning, and model performance tracking, helping boost productivity among the company's developers.

Applied Research Associate at Polytechnique Montreal | Montreal, QC, CA May, 2022 - present

- Lead and coordinate a team of PhDs and postdocs to develop an AI/ML framework, optimizing accuracy and reducing training time in deep neural networks, paving the way for major advancements in various industries.
- Lead the technical development of C++/CUDA/Python open-source software for Bayesian neural networks, cuTAGI, improving the robustness of deep neural networks outcome.
- Write and optimize custom C++/CUDA kernels for Linear, CNNs, LSTM, and Transformer architectures.

Machine Learning Engineer at Shearwater Aerospace | Montreal, QC, CA September, 2021 - June, 2022

- Built machine learning-based path planning system to improve UAV flight efficiency.
- Developed an autonomous control system using reinforcement learning for UAVs, enabling drones to fly from departure to destination without human supervision.

Postdoctoral Researcher at Polytechnique Montreal | Montreal, QC, CA November, 2019 - September, 2021

- Developed a theory-based approach for modeling uncertainty in deep neural networks, enhancing reliability.
- Created an open-sourced package BDLM for detecting anomalies in time series, significantly improving detectability for industrial partners.

EDUCATION

Ph.D. in Computer Science for Civil Engineering at Polytechnique Montreal | Montreal, QC, CA October, 2019

PERSONAL PROJECTS

1. **cuTAGI** for Bayesian Neural Networks (2018-present) | tagiml.com. cuTAGI: An open-source Bayesian neural network developed in C++/CUDA. It quantifies uncertainty in deep neural networks for various learning tasks, enhancing output reliability and accuracy.

2. **Large Language Model for Finance** (2023) | github.com/lhnguyen102/llm-finance. This project is only for educational and learning purposes, focusing on guiding developers on how to build an LLM from scratch and then fine-tune it for different tasks.

3. **Transformer Temporal Fusion** (2023) | github.com/lhnguyen102/tft-sgd. Transformer based approach for time series forecasting

PUBLICATIONS

1. Analytically Tractable Hidden-States Inference in Bayesian Neural Networks. Journal-to-conference track. *ICLR*, 2024.

2. Analytically tractable heteroscedastic uncertainty quantification in Bayesian neural networks for regression tasks, Neurocomputing, 2024.

3. Hiking up that HILL with Cogment-Verse: Train & Operate Multi-agent Systems Learning from Humans, AAMAS, 2023.

4. Tractable Approximate Gaussian Inference for Bayesian Neural Networks. *JMLR*, 2021.

5. Analytically Tractable Inference in Neural Networks-An Alternative to Backpropagation, NeurIPS, 2021.