Sébastien Doyez
(438) 367 6580
2100, rue René H2L1T9 Longueuil, QC

doyez.sebastien34090@gmail.com
Linkedin | Portfolio | Github | HuggingFace
Languages: French, English

Machine Learning & Software Engineer specialized in real-time computer vision, edge AI and production ML pipelines. Experienced in taking models from implementation to low-latency deployment using Python, C++, TensorRT, ONNX, and Docker across cloud and embedded targets. Proven track record in model optimization (quantization and pruning), RAG systems and scalable data/ MLOps workflows.

SKILLS

Machine Learning: Deep Learning, Computer Vision, RAG, LLM, CNNs, Vision Transformers, Semantic Segmentation, Object Detection
Programming: **Python, C (embedded)**, **C++,** SQL, Javascript, HTML/CSS
Frameworks: TensorFlow, PyTorch, ONNX, TensorRT, Qt, PyQt, React, tkinter
Tools: Git, SVN, Linux, Docker.
Embedded Systems: STM32 (bare metal), Raspberry Pi, Arduino, ESP32
Embedded Protocols: UART, I2C, SPI, PWM, FreeRTOS.
Systems & Deployment: Embedded ML, Edge AI, NVIDIA Jetson, Model Optimization, Inference Benchmarking

WORK EXPERIENCES:

Machine Learning / Software Engineer

Servo Robot, Saint-Bruno-de-Montarville, QC, Canada
October 2024 – October 2026

- **High Performance Vision Pipelines:** Achieved real-time inference (150 ms) for multi-class semantic segmentation (welds, joints, holes) using **Swin U-Net, SegFormer, U-Net, U-Net++** and **Attention U-Net** reaching up to 98% IoU across benchmarks.
- **Generative AI & Semantic Search:** Architected and deployed an end-to-end **RAG** system using Vector Databases (Qdrant) to enable semantic querying of complex technical documentation, improving client support efficiency.
- **Scalable ML infrastructure & SQL: Designed** and **maintained 2 database architectures on Cloud VPS**: engineered a **Model Registry** to track fine-tuning iterations per client and a **Data Lake** for structured storage and label versioning, ensuring full data lineage.
- **Production Deployment and MLOps: Successfully deployed** and maintained **7 industrial-grade ML systems in production**, implemented a continuous MLOps workflow for model monitoring and iterative performance improvement with client feedback.
- **Deep Learning Optimization and Compilation:** Designed and deployed industrial-grade vision systems on NVIDIA Jetson Nano (**CUDA**), optimized model runtimes for GPU architecture by converting TensorFlow models to ONNX, compiling them into high performance using TensorRT and implementing TensorFlow Lite (Pruning and Quantization) to reduce memory footprint.
- **Advanced System Tooling:** Developed high-performance **3D visualization software** in **C++20 (Qt/OpenGL)** to display inspection results and integrated interactive custom graphs using **QChart**. Built a custom internal data management and annotation suite using **Python (OOP) and Tkinter**, streamlining the end-to-end data pipeline.
- **Deep Learning on Point Clouds:** Implemented and trained a **PointPillars** model on 3D object detection, optimized the model to extract high-precision edge and corner features while running on NPU i.MX8 (2.5 TOPS).
- **Leadership:** Mentored junior engineers and interns on Machine Learning best practices and **software architecture**.

Robotic Software Engineer Intern,

Sami Agtech, Varennes, QC, Canada
March 2024 – September 2024

- **High- Reliability Backend (C# & Rust)** : Designed and developed mission-critical **microservices**, leveraging its memory safety and concurrency model to build the commercial version of the platform SAMI.
- **Real-Time Computer Vision:** Assisted in training and optimization of deep learning models for real-time object detection (broccoli) within high-bandwidth video streaming pipelines
- **Distributed Systems & Telemetry:** Integrated **MQTT communication** and Fanuc SNPX protocols to synchronize a data flow across a complex robotic cell of 4 robotic Fanuc arms, ensuring seamless communication between hardware and control software.
- **Multithreaded Architecture:** Architected an automatic robotic arm calibration system using **C# (OOP/Multithreading)** and KAREL programming to ensure precise spatial synchronization between sensors and actuators.
- **Low-Level Control:** Programmed complex cutting sequences on **Omron PLCs** using ladder logic, managing real-time state machines and synchronized hardware triggers.

Machine Learning / IoT Engineer Intern, Ets Montréal, Montréal, QC, Canada
June 2023 – August 2023

- **Real-Time Video Telemetry:** Architected a high-performance video streaming and inference pipeline using **NGINX, RTMP, and TCP/IP**, enabling seamless data transfer from a DJI Mavic drone and DJI EP Rover to an embedded edge device.
- **Hardware-Constrained Optimization: Deployed** and **optimized** an EfficientNet-Lite0 CNN for aerial object detection using TensorFlow Lite, achieving **90% IoU** on a Raspberry Pi 3B, reaching the hardware's peak performance **without external accelerators**
- **Advanced Model Compression:** Implemented **pruning** and **quantization** techniques to achieve real-time embedded inference (~2 FPS) on CPU-only hardware, overcoming significant memory and processing constraints.
- **End-to-End Edge Pipeline: I**ntegrated the entire system in **Python,** from raw sensor data acquisition to on-device inference, ensuring reliable performance in a mobile robotics environment.
- **Sensor-Based Positioning:** Developed a secondary vision model for a rover-mounted camera, implementing vision-based positioning algorithms to enhance object grasping accuracy and state estimation.

Robotic Engineer Intern Aide Automatisme, St Quentin, France
June 2022 – July 2022

- **Industrial Automation :** Programmed a robotic arm for perfume cap screwing using Structured Text and Ladder logic within the Beckhoff TwinCAT environment. Used MACtalk software to tune motor speed and acceleration parameters, selecting numerical values that ensured stable operation without oscillations or instabilities.

EDUCATION

Master of Engineering (M.Eng) - Electronics and Computer Science, **Ecole Polytechnique universitaire de Montpellier (Polytech)**
Specialization Robotics, Computer vision and Machine Learning **September 2020 – June 2024**

French preparatory classes for engineering schools (CPGE) **Albert Châtelet, Douai, France**
(equivalent Bachelor's degree) **September 2018, June 2020**

CERTIFICATIONS

AWS Educate Introduction to Cloud 101, Training Badge, **AWS, 2025**
AWS Educate Machine Learning Foundations, Training Badge, **AWS, 2025**
Fundamentals of MCP, **Hugging Face, 2025**
MCP Course: Automation in Production, **Hugging Face, 2025**
AI Agent Fundamental, Certificate of Excellence, **Hugging Face, 2025**
AI Agents Fundamentals, **Hugging Face, 2025**

PERSONAL PROJECTS:

OpenMLCore & MLGUI : Machine Learning Framework & MLOps Interface | Montréal, Canada,

- **Developed `OpenMLCore`,** a **PyTorch** wrapper framework that streamlines deep learning workflows and provides built-in access to pretrained models for **2D** and **3D semantic segmentation.**
- **Architected and built `MLGUI`,** a **PyQt desktop application covering the end-to-end MLOps pipeline** (data labeling, model training, and validation) using a View-Model-Backend pattern to decouple UI state from execution logic and interface directly with `OpenMLCore`.
- **Trained and evaluated 2D segmentation architectures :** achieving up to 0.9802 accuracy, 0.8668 IoU, and 0.9239 F1-score on cell segmentation (U-Net, Attention U-Net) alongside benchmarks on the Dubai dataset (U-Net++), as well as 3D point cloud models reaching 0.9121 (PointNet) and 0.9322 accuracy (PointNet++), with all weights published on HuggingFace.
- Created two dedicated **documentation websites** and produced a YouTube **video speech** showcasing the architecture and goals of both projects.
- **Packaged** and **published** the `OpenMLCore` library on **PyPI for open-source distribution.**

Advanced RAG Agent, Montréal, Canada,

- **Autonomous Agent Design** : Architected a sophisticated AI agent using Python, LangGraph, and QdrantDB, utilizing Gemma (Ollama) for local LLM inference. Implemented a logic-based decision tree to handle complex user queries.
- **Self Correcting Retrieval :** Developed a « Fallback & Search » mechanism : the agent dynamically evaluates RAG confidence scores : if the retrieval probability falls below 50%, it triggers an automated web search (DuckDuckGo) to synthesize up-to-date answers.