Embedded AI Engineer

Company Information

- **Company:** CÔNG TY TNHH MOTOROLA SOLUTIONS VIETNAM R&D CENTER
- **Website:** https://www.motorolasolutions.com/en_us.html
- **Address:** L07.01, Tầng 07, Tháp A, Số 15, Đường Trần Bạch Đằng, Phường An Khánh, Thành phố Hồ Chí Minh, Việt Nam

Job Details

- **Employment Type:** Full-time
- **Work Model:** On-site

Job Description We are seeking a skilled and experienced Embedded AI Engineer to join our team. The successful candidate will be responsible for the development and integration of AI models across diverse hardware platforms, with a strong focus on optimizing for edge device deployment. This role is responsible for delivering high-quality, scalable, and robust AI solutions.

Key Responsibilities Your responsibilities will cover the entire lifecycle of an AI model, from receiving it from the research team to ensuring its efficient operation on the end device.

- **AI Model Optimization & Deployment:** Analyze trained AI models (from TensorFlow, PyTorch) to identify performance bottlenecks and resource requirements (memory, compute). Convert models into inference-optimized formats such as ONNX, TensorRT, and TFLite. Apply advanced optimization techniques, including quantization (INT8/FP16) and pruning, to reduce model size and accelerate processing speed while maintaining accuracy. Profile and benchmark model performance on target hardware (e.g., NVIDIA Jetson, ARM CPUs) to ensure latency and throughput criteria are met.
- **Application Software Development:** Build high-performance applications and libraries in C++/Python to load, manage, and execute AI models in both Linux and Windows environments. Develop end-to-end data processing pipelines, from pre-processing input data (images, video) to post-processing model outputs. Create and maintain unit and integration tests.

Technical Requirements & Qualifications

- **Hardware/Edge Deployment:** Experience deploying AI models on Edge Computing platforms from major vendors such as NVIDIA (Jetson series) or Ambarella (CVflow architecture), utilizing optimization tools like TensorRT, TFlite, and ONNX.
- **OS:** Solid experience in developing and deploying applications on Linux (Ubuntu/Embedded) and Windows operating systems.
- **Deep Learning Frameworks:** Strong proficiency in PyTorch or TensorFlow, with the ability to analyze model graphs, layers, and operations for optimization purposes.
- **Optimization Techniques:** In-depth knowledge of model compression techniques, specifically Quantization (INT8/FP16) and Pruning. Experience with Post-Training Quantization (PTQ) or Quantization-Aware Training (QAT) is a plus.
- **Hardware Profiling:** Experience profiling and benchmarking AI models to identify bottlenecks (latency/throughput) using tools like NVIDIA Nsight Systems, Nsight Compute, trtexec, or Jtop.
- **Computer Architecture:** Solid understanding of memory management, GPU/CPU parallelism, and hardware constraints on embedded systems.
- **Language:** Good command of the English language, both written and verbal.

Preferred Qualifications

- Relevant academic coursework, significant projects, or internships in Artificial Intelligence/Machine Learning or hardware optimization.
- Active participation in technology communities, open-source contributions, or successful hackathon experiences.
- Graduates from highly reputable universities such as HCMUS or HCMUT are particularly encouraged to apply.

Contact:

1.Mr Đinh Phong Quang
E-mail: quang.dinh@motorolasolutions.com
phone/zalo: 0912756999
2. Mr. Lý Quốc Ngọc
E-mail: lqngoc@fit.hcmus.edu.vn