UNTETHER AI

runAI devices

imAIgine software development kit

Key Features

- TensorFlow or PyTorch integration
- Optimized quantization methods
- Post-quantization retraining and knowledge distillation
- Automatic graph lowering and kernel mapping
- Power, area, and performance constraints
- Multi-chip partitioning for large neural networks
- Visualizations for placement and cycle accurate simulation
- Easily integrated runtime API

Overview

The imAIgine Software Development Kit (SDK) enables a push button flow from deep learning pilot model to performant inference implementation on tsunAImi® accelerator cards and runAI200 devices. The imAIgine SDK achieves this by tightly integrating with TensorFlow, PyTorch, and ONNX, enabling custom neural networks to be quantized and optimized for inference. That is backed by a powerful, flexible and easy to integrate client API to get applications up and running quickly.

Applications

The runAI200 devices are designed to accelerate a multiplicity of AI workloads, such as vision-based convolutional networks, RNNs or attention networks for natural language processing, and time-series analysis for financial applications.

Markets	Application	Networks
Vision	Classification, object detection, semantic segmentation	ResNets, YOLO, SSD, Unets, Pose
Natural language processing	Text-to-speech, speech-to-text, chatbots	RNNs, Attention, BERT
Financial technology	X-Value adjustments, credit risk, portfolio balancing	TCNs, LSTMs

imAIgine Software Development Kit

The imAIgine SDK gives developers powerful automated tools and supporting software to quickly go from pilot model to production. It is organized into three parts.

The imAIgine Compiler

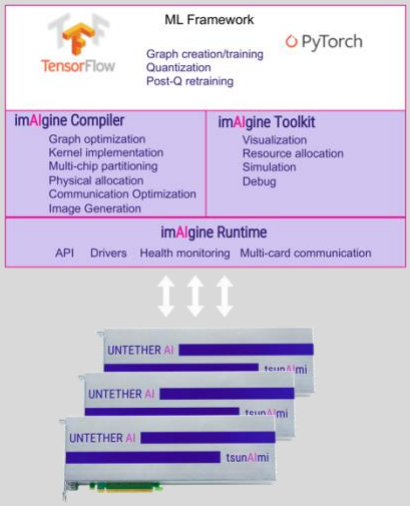
- Import TensorFlow, PyTorch, or ONNX graphs directly
- Automated quantizer and extracts performance without sacrificing accuracy
- Specify performance levels, silicon utilization, and power consumption targets

The imAIgine Toolkit

- Evaluate functionality and performance using the extensive profiling and simulation tools

The imAIgine Runtime

- Provides C-based API for integration into your deep learning environment
- Monitor the health and temperature of the tsunAImi® acceleration cards to ensure proper operation and prevent thermal damage



Familiar frameworks
Quantization and layer optimization done in familiar ML framework

Automated graph lowering
Optimization and allocation algorithms

Extensive feedback
Resource allocations, congestions, cycle-accurate simulation

Easily integrated runtime
Hardware abstraction, communication, and monitoring

Quantizer

Framework Support	TensorFlow, PyTorch, ONNX
Quantization modes	INT8 and INT16, symmetrical and asymmetrical
Quantization-aware retraining	Knowledge Distillation, Labeled QAT, Labeled UAT*

imAIgine Compiler

Kernel Mapping	Untether AI Kernel Library
Custom Kernels	C-based API for kernel development
Spatial Optimizations	Multi-chip partitioning, multi-network per chip, kernel merge

imAIgine Tool Kit

Graph Explorer	Visualize the graph on-chip
Simulator	Cycle-accurate simulations

imAIgine Runtime

Application deployment	Per-chip optimized application deployment
Slice-based Streaming Inference API	Begin inference on streaming data sources as they arrive

*UAT is QAT with a proprietary tuned quantization model

UNTETHER AI