

# imAIne

## 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 imAIne Software Development Kit (SDK) enables a push button flow from deep learning pilot model to performant inference implementation on tsunAI mi<sup>®</sup> accelerator cards and runAI200 devices. The imAIne 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

## imAIne Software Development Kit

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

### The imAIne Compiler

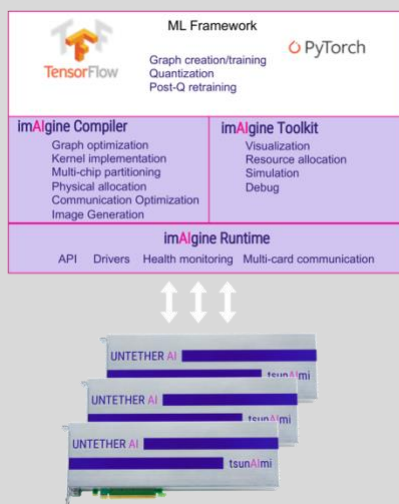
- Import TensorFlow, PyTorch, or ONNX graphs directly
- Automated quantizer and extracts performance without sacrificing accuracy
- Specify performance levels, silicon utilization, and power consumption targets

### The imAIne Toolkit

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

### The imAIne Runtime

- Provides C-based API for integration into your deep learning environment
- Monitor the health and temperature of the tsunAI mi<sup>®</sup> 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*

## imAligne 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

## imAligne Tool Kit

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

## imAligne Runtime

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

### Notice

THE INFORMATION DISCLOSED TO YOU HEREIN (THE "MATERIALS") IS PROVIDED SOLELY FOR THE SELECTION AND USE OF UNTETHER AI'S PRODUCTS. TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, MATERIALS ARE MADE AVAILABLE "AS IS". UNTETHER AI MAKES NO REPRESENTATIONS OR WARRANTIES, WHATSOEVER WITH RESPECT TO THE MATERIALS OR THE PRODUCTS, INCLUDING BUT NOT LIMITED TO REPRESENTATIONS OR WARRANTIES OF MERCHANTABILITY; SECURITY; RELIABILITY; ACCURACY; QUALITY; INTEGRATION; FITNESS FOR A PARTICULAR PURPOSE, NON-INFRINGEMENT, OR TITLE; THAT THE INFORMATION PROVIDED IN THIS MATERIAL IS SUITABLE FOR ANY PURPOSE; NOR THAT THE IMPLEMENTATION OF SUCH INFORMATION WILL NOT INFRINGE ANY THIRD PARTY PATENTS, COPYRIGHTS, TRADEMARKS, OR OTHER RIGHTS. WITHOUT LIMITING THE GENERALITY OF THE FOREGOING, UNTETHER AI EXPRESSLY DISCLAIMS ANY REPRESENTATION, CONDITION, OR WARRANTY THAT ANY INFORMATION PROVIDED TO YOU HEREUNDER, CAN OR SHOULD BE RELIED UPON BY YOU FOR ANY PURPOSE WHATSOEVER. UNTETHER AI DISCLAIMS ANY AND ALL LIABILITY RELATED TO THIS MATERIAL AND WILL NOT BE LIABLE FOR ANY LOSSES OR DAMAGE CAUSED BY RELIANCE ON THE INFORMATION IN THIS MATERIAL.

No license, either expressed or implied, is granted for any intellectual property rights of Untether AI or any third party through the information in this Material. Untether AI shall not be liable (whether in contract or tort, including negligence, or under any other theory of liability) for any loss or damage of any kind or nature related to, arising under, or in connection with, the Materials (including your use of the Materials), including for any direct, indirect, special, incidental, or consequential loss or damage (including loss of data, profits, goodwill, or any type of loss or damage suffered as a result of any action brought by a third party) even if such damage or loss was reasonably foreseeable or Untether AI had been advised of the possibility of the same. Untether AI assumes no obligation to correct any errors contained in the Materials or to notify you of updates to the Materials or to any products. You may not reproduce, modify, distribute, or publicly display the Materials without Untether AI's prior written consent. You should obtain the latest relevant Material before placing orders and should verify that such information is current and complete. All orders are subject to Untether AI's contract which outlines any applicable terms and conditions for a product.

\*UAT is QAT with a proprietary tuned quantization model

### Trademarks

Untether AI, tsunAlmi, runAI, imAligne SDK are trademarks and/or registered trademarks of Untether AI Corporation in the U.S and other countries. Other company names may be trademarks of the respective companies.

### Copyright

© 2023 Untether AI Corporation. All rights reserved.

**UNTETHER AI**