Energy-efficient TCN-Extensions for a TNN accelerator

Tim Fischer
fischeti@iis.ee.ethz.ch

Special Thanks to:
Moritz Scherer, Georg Rutishauser, Matteo Spallanzani, Luca Benini
Processing temporal data on the edge

- Temporal data is very prevalent in edge applications
  - Audio (speech recognition)
  - Biomedical signals (EEG, ECG)
  - Video (gesture recognition)

- Edge resources are scarce
  - Battery powered: ~mWs vs. MFLOPS
  - Small memory: 100s KBs vs. Mparams

- Processing time-series data is complex
  - Additional dimension
  - Capture temporal dependencies
CUTIE – Ternary NN accelerator

What is the SoA for energy-efficient processing on the edge?

- Ternary quantization
  - Encodes 3 values {-1, 0, 1}
  - Achieves reasonable accuracy for some applications
  - Better than binary accuracy
  - Higher efficiency due to sparsity

- **CUTIE – Completely Unrolled Ternary Inference Engine** [1]
  - Hardware accelerator for ternary CNNs
  - Peak energy efficiency: 3.1 [7nm] resp. 0.6 [22nm] POp/s/W @ 0.65V

Time-series processing on the edge

- **Steps required for time-series processing**
  - Feature extraction to generate embeddings from frames (i.e. CNN)
  - Capturing temporal dependencies (i.e. RNN)
  - Classify output (i.e. FC)
Contribution – Ternary TCNs

How can we bring time-series data processing to the edge?

- Temporal Convolutional Neural Networks (TCNs)
  - Causal 1D Convolution
  - Flexible receptive field
  - High parallelism and low memory requirements
  - Competitive accuracy w.r.t. RNNs

- Contributions
  - Ternarization of TCNs
  - Mapping of 1D TCN layers to 2D CNN layers
  - Reuse existing hardware of CUTIE
  - Leverage high parallelism of CNN layers
TNN-Accelerator - CUTIE

Completely Unrolled Ternary Inference Engine

- Minimize Data movement
  - Local memory for weights and activations
  - Maximize data reusability
TNN-Accelerator - CUTIE

**Completely Unrolled**

- Minimize Data movement
  - Local memory for weights and activations
  - Maximize data reusability

- Maximize computation efficiency
  - Fully unrolled inner product
  - One compute unit per channel
TNN-Accelerator - CUTIE

Completely Unrolled

- Minimize Data movement
  - Local memory for weights and activations
  - Maximize data reusability

- Maximize computation efficiency
  - Fully unrolled inner product
  - One compute unit per channel

864 ternary multipliers per OCU
82,944 MAC operations in one cycle!!
CUTIE – TNN Accelerator

Completely Unrolled Ternary Inference Engine

- Minimize Data movement
  - Local memory for weights and activations
  - Maximize data reusability

- Maximize computation efficiency
  - Fully unrolled inner product
  - One compute unit per channel

- Minimize switching activity
  - Exploit sparsity of values
Mapping - TCN to CNN

How can we map 1D TCNs to 2D CNNs?

Length = 8
Dilation = 4

Width = 4
Height = 4
Mapping – TCN to CNN

Width = 8
Dilation = 1

Width = 1
Height = 10
Mapping – TCN to CNN

Width = 8
Dilation = 4

Width = 4
Height = 4
Implementation - TCN

- Control is extended to support TCNs
  - Very minimal hardware modifications
  - Mapping is performed in software
- TCN activation memory

Control is extended to support TCNs
- Very minimal hardware modifications
- Mapping is performed in software

TCN activation memory
What is the cost for TCN support?

- **Area overhead**
  - Control modifications
  - Additional activation memory

- **GF 22 nm technology**
  - 96 input/output channels, 64x64 feature maps
  - 8-layer, 3x3 kernels
  - 1.86 mm² / 9.3MGE, post-synthesis
  - Only 1% increase for TCN support
Results – Power and Energy

- **GF 22 nm technology**
  - 0.65 V @ 66 MHz, post-synthesis
  - Peak energy efficiency: 910 TOp/s/W
  - Avg. energy efficiency: 450 TOp/s/W

- **Inference on DVSGesture**
  - 92.6% accuracy
  - Core energy per inference: 1 µJ
  - Core throughput: 5.48 TOp/s
  - Core inference Time: 78 µs
Conclusion

- Processing temporal data on the edge is possible
  - Ternary TCNs are well-suited for edge processing

- Smart Mapping of 1D TCN to 2D CNN
  - Mapping is primarily done in software
  - only 1% hardware area increase

- CUTIE achieves exceptional energy-efficiency
  - Peak performance of 910 TOp/s/W in GF 22 nm
Premier Sponsor
Automated TinyML

Zero-code SaaS solution

Create tiny models, ready for embedding, in just a few clicks!

Compare the benchmarks of our compact models to those of TensorFlow and other leading neural network frameworks.

Executive Sponsors
Optimized models for embedded
Application
Optimized low-level NN libraries
(i.e. CMSIS-NN)
Runtime
(e.g. TensorFlow Lite Micro)
RTOS such as Mbed OS
Arm Cortex-M CPUs and microNPUs
Connect to high-level frameworks
Profiling and debugging tooling such as Arm Keil MDK
Supported by end-to-end tooling
Connect to Runtime

Stay Connected

Resources: developer.arm.com/solutions/machine-learning-on-arm
TinyML for all developers

Acquire valuable training data securely

Enrich data and train ML algorithms

Edge Device
Real sensors in real time
Open source SDK

Embedded and edge compute deployment options

Test impulse with real-time device data flows

www.edgeimpulse.com
Advancing AI research to make efficient AI ubiquitous

Power efficiency
- Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization
- Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning
- Robust learning through minimal data, unsupervised learning, on-device learning

Perception
- Object detection, speech recognition, contextual fusion

Reasoning
- Scene understanding, language understanding, behavior prediction

Action
- Reinforcement learning for decision making

A platform to scale AI across the industry

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
Syntiant Corp. is moving artificial intelligence and machine learning from the cloud to edge devices. Syntiant’s chip solutions merge deep learning with semiconductor design to produce ultra-low-power, high performance, deep neural network processors. These network processors enable always-on applications in battery-powered devices, such as smartphones, smart speakers, earbuds, hearing aids, and laptops. Syntiant's Neural Decision Processors™ offer wake word, command word, and event detection in a chip for always-on voice and sensor applications.

Founded in 2017 and headquartered in Irvine, California, the company is backed by Amazon, Applied Materials, Atlantic Bridge Capital, Bosch, Intel Capital, Microsoft, Motorola, and others. Syntiant was recently named a CES® 2021 Best of Innovation Awards Honoree, shipped over 10M units worldwide, and unveiled the NDP120 part of the NDP10x family of inference engines for low-power applications.

www.syntiant.com @Syntiantcorp
Platinum Sponsors
Part of your life. Part of tomorrow.

www.infineon.com
Add Advanced Sensing to your Product with Edge AI / TinyML

Pre-built Edge AI sensing modules, plus tools to build your own

Reality AI solutions
- Prebuilt sound recognition models for indoor and outdoor use cases
- Solution for industrial anomaly detection
- Pre-built automotive solution that lets cars “see with sound”

Reality AI Tools® software
- Build prototypes, then turn them into real products
- Explain ML models and relate them to the physics
- Optimize the hardware, including sensor selection and placement

https://reality.ai  info@reality.ai  @SensorAI  Reality AI
Gold Sponsors
Adaptive AI for the Intelligent Edge

LatentAI.com
Build Smart IoT Sensor Devices From Data

SensiML pioneered TinyML software tools that auto generate AI code for the intelligent edge.

- End-to-end AI workflow
- Multi-user auto-labeling of time-series data
- Code transparency and customization at each step in the pipeline

We enable the creation of production-grade smart sensor devices.
Silver Sponsors
Copyright Notice

The presentation(s) in this publication comprise the proceedings of tinyML® EMEA Technical Forum 2021. The content reflects the opinion of the authors and their respective companies. This version of the presentation may differ from the version that was presented at tinyML EMEA. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyML.org