VOICE YOUR FUTURE
Moshe Haiut is a principle senior engineer in the CTO team at DSP Group, Herzliya, Israel. His main areas of expertise are communication, video and audio digital signal processing, digital hardware architecture, and neural networks. Moshe holds M.Sc (with honors) and B.Sc. (with honors) degrees in Electrical Engineering from Tel-Aviv University and the Technion, Israel. Moshe is the architect of the nNetLite – an ultra-low-power H/W engine for NN inference models execution.
The Problem – Memory space for the parameters (weights)

• Small-to-medium NN models may have millions of parameters, while the tinyML SOC usually has limited memory (hundreds of KBytes) to store the weights data

• In this presentation we show how this problem was solved by an innovative h/w accelerator (Weights Extraction Unit) combined with a powerful s/w toolchain

Quantization
Pruning
Entropy Coding
Packing
The DSPG nNetLite h/w Engine

- The nNetLite engine is a stand-alone module, planned to be embedded in DSP Group’s future SOCs.
- The DBM10L ultra-low-power device comprises the nNetLite engine and the CEVA TeakLite-3 side-by-side.
### Weights Compression by the nNetLite Compiler

#### Asymmetric Quantization

<table>
<thead>
<tr>
<th>16-bit</th>
<th>11-bit</th>
<th>3-bit</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.11290641129016876</td>
<td>0</td>
<td>0.001825837185606360</td>
</tr>
<tr>
<td>0.00341823790224064</td>
<td>0</td>
<td>0.00341823790224064</td>
</tr>
<tr>
<td>0.3417060710489532</td>
<td>0</td>
<td>5.799067206680776</td>
</tr>
<tr>
<td>5.799067206680776</td>
<td>0</td>
<td>0.8514375658705831</td>
</tr>
<tr>
<td>0.8514375658705831</td>
<td>0</td>
<td>1.8651321297511458</td>
</tr>
<tr>
<td>1.8651321297511458</td>
<td>0</td>
<td>4.0</td>
</tr>
<tr>
<td>4.0</td>
<td>0</td>
<td>5.886017650365839</td>
</tr>
<tr>
<td>5.886017650365839</td>
<td>0</td>
<td>0.8896064944565296</td>
</tr>
<tr>
<td>0.8896064944565296</td>
<td>0</td>
<td>0.00341823790224064</td>
</tr>
<tr>
<td>0.00341823790224064</td>
<td>0</td>
<td>0.001825837185606360</td>
</tr>
</tbody>
</table>

#### Pruning

<table>
<thead>
<tr>
<th>11-bit</th>
<th>3-bit</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>011</td>
</tr>
<tr>
<td>001</td>
<td>010</td>
</tr>
<tr>
<td>011</td>
<td>001</td>
</tr>
</tbody>
</table>

#### Entropy Coding

<table>
<thead>
<tr>
<th>3-bit</th>
<th>11-bit</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>010010111</td>
</tr>
<tr>
<td>011</td>
<td>00010011</td>
</tr>
<tr>
<td>000</td>
<td>00010011</td>
</tr>
<tr>
<td>010</td>
<td>01100000</td>
</tr>
<tr>
<td>011</td>
<td>01000000</td>
</tr>
<tr>
<td>001</td>
<td>01110000</td>
</tr>
</tbody>
</table>

#### Packing

<table>
<thead>
<tr>
<th>3-bit</th>
<th>11-bit</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>01011001</td>
</tr>
<tr>
<td>111</td>
<td>01111000</td>
</tr>
<tr>
<td>001</td>
<td>11000000</td>
</tr>
</tbody>
</table>

---

6 | Private and Confidential. Copyright DSP Group, 2018. All rights reserved.
Weights De-compression by the nNetLite WEU h/w module

Unpacking
01011001
01111000

Entropy Decoding
0 10110
0 101
11 1

De-quantization
000 011 000 010 011 001

16-bit Q7.8

Decimal value
0.0000000
5.79687500
0.0000000
1.86328125
5.88671875
0.89062500

0.0000000
5.79687500
0.0000000
1.86328125
5.88671875
0.89062500
Weights De-compression by the nNetLite WEU h/w module

SOC 1MB RAM

NN Compressed Weights

DMA

Weights Extraction Unit (WEU)

De-compressor

De-quantizer

16 KB Cyclic SRAM (8K X 16-bit)

Valid weights

Valid weights

RD pointer

WR pointer

16-bit MAC Unit

16-bit MAC Unit
Example: Face Detection NN

- Number of layers: 8
- Number of MAC ops: ~12M / Inference
- Number of Weights: 224,656
- Available memory: 280KB
w/o Compression. Weights: 439 KB, RMSE: 0.130191
12-bit weights, 50% Prun. Weights: 273 KB, RMSE: 0.123725
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 applications (e.g. TensorFlow Lite Micro)

Optimized low-level NN libraries (i.e. CMSIS-NN)

Arm Cortex-M CPUs and microNPUs

RTOS such as Mbed OS

Profiling and debugging tooling such as Arm Keil MDK

Supported by end-to-end tooling

Connect to Runtime

Connect to high-level frameworks

AI Ecosystem Partners

Stay Connected

@ArmSoftwareDevelopers

@ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm

Arm: The Software and Hardware Foundation for tinyML
TinyML for all developers

Dataset
- 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

Impulse
- 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](http://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 the function 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.

sensiml.com
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