“Always-on visual classification below 1 mW with spiking convolutional networks on Dynap™-CNN”

Martino Sorbaro – SynSense

February 2nd, 2021
Arm: The Software and Hardware Foundation for tinyML

1. Connect to high-level frameworks
   - Profiling and debugging tooling such as Arm Keil MDK

2. Supported by end-to-end tooling
   - Optimized models for embedded
     - Runtime (e.g. TensorFlow Lite Micro)

3. Connect to Runtime
   - Optimized low-level NN libraries (i.e. CMSIS-NN)
   - RTOS such as Mbed OS
   - Arm Cortex-M CPUs and microNPUs

Stay Connected
- @ArmSoftwareDevelopers
- @ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm
WE USE AI TO MAKE OTHER AI FASTER, SMALLER AND MORE POWER EFFICIENT

- **Automatically compress** SOTA models like MobileNet to <200KB with little to no drop in accuracy for inference on resource-limited MCUs
- **Reduce** model optimization trial & error from weeks to days using Deeplite's design space exploration
- **Deploy more** models to your device without sacrificing performance or battery life with our easy-to-use software

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

Test

- Test impulse with real-time device data flows

Get your free account at http://edgeimpulse.com
Health sensors measure PPG and ECG signals critical to understanding vital signs. Signal chain products enable measuring even the most sensitive signals.

The biggest (3MB flash and 1MB SRAM) and the smallest (256KB flash and 96KB SRAM) Cortex M4 microcontrollers enable algorithms and neural networks to run at wearable power levels.

The new MAX78000 implements AI inferences at over 100x lower energy than other embedded options. Now the edge can see and hear like never before.
Qeexo AutoML for Embedded AI
Automated Machine Learning Platform that builds tinyML solutions for the Edge using sensor data

Key Features

- Wide range of ML methods: GBM, XGBoost, Random Forest, Logistic Regression, Decision Tree, SVM, CNN, RNN, CRNN, ANN, Local Outlier Factor, and Isolation Forest
- Easy-to-use interface for labeling, recording, validating, and visualizing time-series sensor data
- On-device inference optimized for low latency, low power consumption, and a small memory footprint
- Supports Arm® Cortex™- M0 to M4 class MCUs
- Automates complex and labor-intensive processes of a typical ML workflow – no coding or ML expertise required!

Target Markets/Applications

- Industrial Predictive Maintenance
- Smart Home
- Wearables
- Automotive
- Mobile
- IoT

For a limited time, sign up to use Qeexo AutoML at automl.qeexo.com for FREE to bring intelligence to your devices!
is for building products

Reality AI Tools® software

- Automated Feature Exploration and Model Generation
- Bill-of-Materials Optimization
- Automated Data Assessment
- Edge AI / TinyML code for the smallest MCUs

Reality AI solutions

- Automotive sound recognition & localization
- Indoor/outdoor sound event recognition
- RealityCheck™ voice anti-spoofing

https://reality.ai  info@reality.ai  @SensorAI  Reality AI
SynSense builds ultra-low-power (sub-mW) sensing and inference hardware for embedded, mobile and edge devices. We design systems for real-time always-on smart sensing, for audio, vision, IMUs, bio-signals and more.

https://SynSense.ai
Announcement

www.tinyML.org/summit2021

Highlights:
- Keywords: Premier Quality, Interactive, LIVE ... and FREE
- 5 days, 50+ presentations
- 4 Tutorials
- 2 Panel discussions: (i) VC and (ii) tinyML toolchains
- tinyML Research Symposium
- Late Breaking News
- 3 Best tinyML Awards (Paper, Product, Innovation)
- 10+ Breakout sessions on various topics
- tinyML Partner sessions
- tinyAI for (Good) Life
- LIVE coverage, starting at 8am Pacific time

What should I do about it:
- Check out the program – you will be impressed
- Register on-line (takes 5 min, and 1000+ already registered in 5 days)
- If interested: Submit nominations for Best Awards and/or Late News – February 28 deadline
- Block out your calendar: March 22-26
- Become a sponsor (sponsorships@tinyML.org)
- Spread out the word about the Summit
- Actively participate at the Summit
- Provide your feedback – we listen!
- Don’t worry about missing some talks – all videos will be posted on YouTube.com/tinyML
tinyML is growing fast

<table>
<thead>
<tr>
<th></th>
<th>2019 Summit (March 2019)</th>
<th>2020 Summit (Feb 2020)</th>
<th>2021 Summit (March 2021), expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attendees</td>
<td>160</td>
<td>400+</td>
<td>3000+</td>
</tr>
<tr>
<td>Companies</td>
<td>90</td>
<td>172</td>
<td></td>
</tr>
<tr>
<td>Linkedin members</td>
<td>0</td>
<td>798</td>
<td>~ 2000</td>
</tr>
<tr>
<td>Meetups members</td>
<td>0</td>
<td>1140</td>
<td>~ 5000</td>
</tr>
<tr>
<td>YouTube subscribers</td>
<td>0</td>
<td>0</td>
<td>~ 3000</td>
</tr>
</tbody>
</table>

also started in Asia: tinyML WeChat and BiliBili
Summit Sponsors
(as of Feb 1, 2021)

Contact: sponsorships@tinyML.org

multiple levels and benefits available
(also check www.tinyML.org)
Next tinyML Talks

<table>
<thead>
<tr>
<th>Date</th>
<th>Presenter</th>
<th>Topic / Title</th>
</tr>
</thead>
</table>
| Tuesday, February 16 | **Mohammed Zubair**  
PhD, SMIEEE  
Associate Professor / Consultant  
Department of Electrical Engineering /  
Center for Artificial Intelligence  
King Khalid University | Oral Tongue Lesion Detection using TinyML on Embedded Devices |

Webcast start time is 8 am Pacific time

Please contact talks@tinyml.org if you are interested in presenting
Reminders

Slides & Videos will be posted tomorrow

Please use the Q&A window for your questions

tinyml.org/forums  youtube.com/tinyml
Martino Sorbaro

Martino Sorbaro is a research and development scientist at SynSense AG, Zürich, Switzerland, and a postdoc at the institute of Neuroinformatics of the University of Zürich and ETH. He obtained a MSc in physics at the university of Pavia, Italy, and a PhD in neuroinformatics at the university of Edinburgh, Scotland. His current work focuses on learning in spiking neural networks, both for theoretical research and technological applications.
Always-on visual classification below 1 mW with spiking convolutional networks on Dynap-CNN

Martino Sorbaro

SynSense
Institute of Neuroinformatics
(UZH/ETH)
What is Neuromorphic Engineering?

Standard **deep learning**

• Runs on PC + CPU/GPU/TPU (or specialized accelerators)
• Continuous values
• Synchronous

**Neuromorphic models**

• Runs on specialized hardware
• Discrete signals (“spikes”)*
• Asynchronous, continuous time
What is Neuromorphic Engineering?

**Standard deep learning**

\[ y = \Theta(W \cdot x + b) \]

Inputs: \( x \)

Output: \( y \)

Weights: \( W \)

Bias: \( b \)

**Neuromorphic models**

\[ z_o(t) = \Theta[V_m(t)] \]

Output (events): \( z_o(t) \)

Transfer function: \( \Theta \)

Internal state: \( V_m(t) \)

Synaptic inputs: \( l(t) \)

Neuron state: \( V_m(t) \)

Input spikes: \( z(t) \)

Bias input: \( b \)
Application domains

**Ultra-low power natural signal processing (Spiking NN)**

- Embedded healthcare
- Always-on audio processing
- Industrial monitoring

**High speed, low power vision processing (Spiking CNN)**

- Visual detection, tracking, recognition for autonomous vehicles and drones
- Smart presence and behaviour detection
- Real-time visual motion estimation
Dynap-CNN

A **convolutional** neuromorphic chip

- 9 convolutional layers
- Up to 1024 feature channels
- 16 output classes

- **Vision** applications
Digital spiking neuron

Input synaptic events

1 bit

Synaptic weights

8 bits
Weight

8 bits
Weight

Threshold

16 bits
Neuron state

Output event

1 bit
Event-driven vision

- **Dynamic Vision Sensors** — Neuromorphic asynchronous cameras
  - Very high dynamic range
  - No frames, instantaneous signal from each pixel
  - No power expended if nothing moves
Event-driven vision processing

CNN-based processing

Input array → Clipping → Convolutional layer → Pooling → Dense layer → Pooling

- 128x128 input array
- Clipping
- Convolutional layer: Kernel up to 16x16, Feature up to 64x64, Configurable stride
- Pooling
- Dense layer: Up to 1024 neurons

Up to 9 convolutional/dense layers, Up to 278k parameters and 327k neurons total
Training Spiking CNNs

Deep artificial network

The brain’s visual cortex

Backpropagation of error

neuroconllab.org
Training Spiking CNNs

In the brain

- Biologically plausible learning
  (open area of research)

Artificial spiking network

- Backprop through time
  (slow, memory demanding)

- Weight transfer from normal CNN
  (fast, no patterns in time)
Training Spiking CNNs

DVS Recording

Spikes accumulated into static frames

Use same weights in SNN (rate coding)

Train CNN with backprop
Optimizing energy consumption

Energy consumption $\sim$ number of “synaptic operations”

$\text{Loss} = \text{Error} + \text{#SynOps}$
Optimizing energy consumption

Quantized ReLU

• Prevents silencing of activity

• Requires surrogate gradient
Optimizing energy consumption

Sorbaro, Liu, Bortone, and Sheik, Frontiers in Neuroscience, 2020

Spiking MNIST dataset

Accuracy versus activity, spiking network

SynOps by layer

Accuracy

Sorbaro, Liu, Bortone, and Sheik, Frontiers in Neuroscience, 2020
Optimizing energy consumption

CIFAR-10 dataset

Per inference \(~1\) mJ at \(~90\)% accuracy

Example: gesture recognition

IBM Gestures dataset

Our state of the art:
• 10 classes
• ~89% accuracy (1 s gestures)
• 0.33 mW estimated power use
Comparisons

Comparisons

Caveats

• Neuromorphic power use changes with input
• Different network sizes supported
• Power per inference vs. per second
• Accuracy per inference vs. per second
  • DynapCNN 1 CIFAR-10 inference: \( \sim 1 \text{ mJ at } \sim 90\% \text{ accuracy} \)
  • Ambarella automotive camera: 3W at 30 fps = 100 mJ (higher res.)
Example: gesture recognition
Example: face presence detection
Future

• A chip for signal processing in time for low-dimensional signals (audio, bio-signals)

• New application scenarios

• Dynap-CNN development kit currently available in limited amounts
DynapCNN development kit
Thanks for listening!

martino.sorbaro@synsense.ai
http://synsense.ai
This presentation in this publication was presented as a tinyML® Talks webcast. The content reflects the opinion of the author(s) and their respective companies. 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