On Device Learning Forum

Enabling Ultra-low Power Machine Learning at the Edge

“On-Chip Learning and Implementation Challenges with Oscillatory Neural Networks”

Aida Todri-Sanial – Professor, Eindhoven University of Technology

May 16, 2023

www.tinyML.org
The goal of On Device Learning (ODL) is to make edge devices “smarter” and more efficient by observing changes in the data collected and self-adjusting/reconfiguring the device’s operating model. Optionally the “knowledge” gained by the device is shared with other deployed devices.

Danilo Pau, Elias Fallon, Evgeni Gousev, Davis Sawyer, Ira Feldman, Christopher B. Rogers
tinyML On Device Learning Forum
8/31 – 9/1, 2022 Online

Accademia on 8/31/2022
– On-Device Learning Under 256KB Memory, Song HAN, Assistant Professor, MIT EECS
– Neural Network ODL for Wireless Sensor Nodes, Hiroki MATSUTANI, Professor, Keio University
– Scalable, Heterogeneity-Aware and Trustworthy Federated Learning, Yiran CHEN, Professor, Duke University
– On-Device Learning For Natural Language Processing with BERT, Warren J. GROSS, Professor, McGill University
– Is on-device learning the next “big thing” in TinyML? Manuel ROVERI, Associate Professor, Politecnico di Milano
– ODL Professors Panel

Industry on 9/1/2022
– TinyML ODL in industrial IoT, Haoyu REN, PhD Student, Technical University of Munich/Siemens
– NeuroMem® wearable, hardwired sub milliwatt real time machine learning with wholly parallel access to “neuron memories” fully explainable, Guy PAILLET, Co-founder, General Vision
– Using Coral Dev Board Micro for ODL innovations, Bill LUAN, Senior Program Manager, Google
– Platform for Next Generation Analog AI Hardware Acceleration, Kaoutar EL MAGHRAOUI, Principal Research Scientist, IBM T.J Watson Research Center
– Enabling on-device learning at scale, Joseph SORIAGA, Sr. Director of Technology, Qualcomm
– Training models on tiny edge devices, Valeria TOMASELLI, Senior Engineer, STMicroelectronics

https://www.tinyml.org/event/on-device-learning/
On device learning Forum

- A framework of algorithms and associated tool for on-device tiny learning, Danilo PAU, Technical Director, IEEE and ST Fellow, STMicroelectronics

- In Sensor and On-device Tiny Learning for Next Generation of Smart Sensors Michele MAGNO, Head of the Project-based learning Center, ETH Zurich, D-ITET

- Continual On-device Learning on Multi-Core RISC-V MicroControllers Manuele RUSCI, Embedded Machine Learning Engineer, Greenwaves

- On-device continuous event-driven deep learning to avoid model drift, Bijan MOHAMMADI, CSO, Bondzai

https://www.tinyml.org/event/on-device-learning/
2023 Gartner Emerging Technologies and Trends Impact Radar

gartner.com

Note: Range measures number of years it will take the technology/trend to cross over from early adopter to early majority adoption. Mass indicates how substantial the impact of the technology or trend will be on existing products and markets.

Source: Gartner
© 2023 Gartner, Inc. All rights reserved.
On Device Learning Forum 2023, May 16 2023

- 8:00 - 8:10 Opening remarks by Danilo Pau
- 8:10 - 8:40 Charlotte Frenkel "Merging insights from artificial and biological neural networks for neuromorphic edge intelligence"
- 8:40 - 9:40 Giorgia Della Ferrera "Forward Learning with Top-Down Feedback: Solving the Credit Assignment Problem without a Backward Pass"
- 9:40 - 10:10 Guy Paillet "NeuroMem®, Ultra Low Power hardwired incremental learning and parallel pattern recognition"
- 10:10 - 10:40 Aida Todri-Sanial "On-Chip Learning and Implementation Challenges with Oscillatory Neural Networks"
- 10:40 - 11:10 Eduardo S. Pereira “Online Learning TinyML for Anomaly Detection Based on Extreme Values Theory”
- 11:10 - 11:15 Closing remarks by Danilo Pau
Thank you, tinyML Strategic Partners, for committing to take tinyML to the next Level, together
Executive Strategic Partners
EDGE IMPULSE

The Leading Development Platform for Edge ML

dgeimpulse.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.
Accelerate Your Edge Compute

SYNTIANT

Making Edge AI A Reality

www.syntiant.com
Platinum Strategic Partners
Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.
DEPLOY VISION AI
AT THE EDGE AT SCALE
Gold Strategic Partners
Witness potential made possible at analog.com.
Easily deploy your tinyML solutions with Arduino Pro

arduino.cc/pro

Made In Italy
Arm AI Virtual Tech Talks

The latest in AI trends, technologies & best practices from Arm and our Ecosystem Partners.

Demos, code examples, workshops, panel sessions and much more!

Fortnightly Tuesday @ 4pm GMT/8am PT

Find out more: www.arm.com/techtalks
Driving decarbonization and digitalization. Together.

Infineon serving all target markets as Leader in Power Systems and IoT

www.infineon.com
NEUROMORPHIC INTELLIGENCE FOR THE SENSOR-EDGE
The Right Edge AI Tools Can Make or Break Your Next Smart IoT Product

Analytics Toolkit Suite

AutoML

Data Collection

tool & Validation

Code Generation

Version Control and Model Management

sensiML.com/tinyML
STMicroelectronics provides extensive solutions to make tiny Machine Learning easy
ENGINEERING
EXCEPTIONAL
EXPERIENCES

We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

www.synaptics.com
Join Growing tinyML Communities:

**Meetup**

tinyML - Enabling ultra-low Power ML at the Edge

**LinkedIn**

The tinyML Community
https://www.linkedin.com/groups/13694488/

14.7k members in 47 Groups in 39 Countries

4k members & 11.6k followers
Subscribe to tinyML YouTube Channel for updates and notifications (including this video)
www.youtube.com/tinyML
tinyML EMEA Innovation Forum

June 26 -28, 2023
Amsterdam

EMEA 2023
https://www.tinyml.org/event/emea-2023

More sponsorships are available: sponsorships@tinyml.org
Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums  youtube.com/tinyml

Please use the Q&A window for your questions
Aida Todri-Saniel received the B.S. degree in electrical engineering from Bradley University, IL in 2001, M.S. degree in electrical engineering from Long Beach State University, CA, in 2003 and a Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara, in 2009. She is currently a Full Professor in Electrical Engineering Department at Eindhoven University of Technology, Netherlands and Director of Research for the French National Council of Scientific Research (CNRS). Dr. Todri-Saniel was a visiting fellow at the Cambridge Graphene Center and Wolfson College at the University of Cambridge, UK during 2016-2017. Previously, she was an R&D Engineer for Fermi National Accelerator Laboratory, IL. She has also held visiting research positions at Mentor Graphics, Cadence Design Systems, STMicroelectronics and IBM TJ Watson Research Center. Her research interests focus on emerging technologies and novel computing paradigms such as neuromorphic and quantum computing.
On-Chip Learning and Implementation Challenges with Oscillatory Neural Networks

16 MAY 2023  TINYML ON DEVICE LEARNING FORUM 2023

Madeleine Abernot and Aida Todri-Sanial, NanoComputing Research Lab

Electrical Engineering Department, Electronics System
Oscillatory Neural networks

NEUROMORPHIC COMPUTING FOR AI AT THE EDGE

- Support online learning
- Fast and efficient inference
- Low power consumption
- Scalability
- Low cost

*Image: Kim S., et al., Nanoscale. 2020*
Oscillatory Neural networks

**NEUROMORPHIC COMPUTING FOR AI AT THE EDGE**

- Support online learning
- Fast and efficient inference
- Low power consumption
- Scalability
- Low cost

**PHASE COMPUTING PARADIGM**

- Brain-inspired computing paradigm
- Neurons are oscillators
- Synapses are coupling elements between oscillators
- Information encoded in oscillators’ phases

*Image: Kim S, et al., Nanoscale. 2020*
ONN Computing Paradigm

ONN paradigm\textsuperscript{1,2,3}:
Network of coupled oscillators
Phase computation

\textsuperscript{1}G. Csaba et al., \textit{Applied Physics Review}, 2020.
\textsuperscript{2}J. Shamsi \textit{et al.}, Front. In Neuroscience, 2021.
\textsuperscript{3}C. Delacour \textit{et al.}, \textit{ISVLSI}, 2021.
ONN Computing Paradigm

ONN paradigm\(^1,2,3\):
- Network of coupled oscillators
- Phase computation

State of the art:
- Associative memory
- Unsupervised learning rules introduced for HNN\(^4\)

Digital ONN implementation

ONN on FPGA implementation
Neurons: phase-controlled oscillators
Synapses: n-bits signed registers
Off-chip learning
Fully-connected design
Zybo-Z7 board

ONN on FPGA implementation

- Tested on two ONN sizes (15 / 60 neurons)
- Pattern recognition task (fully-connected)
- 5-bits signed registers for synapses
- Due to resources limitation:
  - maximum size: 100 neurons

<table>
<thead>
<tr>
<th>ONN size</th>
<th>Synapses</th>
<th>FPGA resources</th>
<th>Init. time (us)</th>
<th>Comp. time (us)</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>5x3</td>
<td>225</td>
<td>1,8% 0,68%</td>
<td>2</td>
<td>5,2</td>
<td>141000</td>
</tr>
<tr>
<td>10x6</td>
<td>3600</td>
<td>12% 2,6%</td>
<td>7,8</td>
<td>5,4</td>
<td>75000</td>
</tr>
</tbody>
</table>

Digital ONN for Pattern Recognition

10x6 ONN:
- 5 training patterns: digits from 0 to 4
- Training: Unsupervised learning rule (ex: Diederich Opper I)
Input image from a camera stream
- Noisy images with flipped pixels (10, 20, 30)
Output image printed on a screen
ONN Learning

ONN learning constraints (hardware):
- Symmetric weight matrix
- Zero-diagonal weight matrix

Constraints for on-chip learning
- Local learning rules
- Incremental learning rules

Classification of various learning rules\(^1\) depending on the ONN on-chip learning constraints.

<table>
<thead>
<tr>
<th>Learning rules</th>
<th>Weight symmetry</th>
<th>Zero-diagonal</th>
<th>Local</th>
<th>Incremental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hebbian</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Storkey</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Diederich Opper I</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Diederich Opper II</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Gardner</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Krauth Mezard</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Pseudo-Inverse</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

\(^1\)P. Tolmachev et al., IJCNN, 2020.

\[ \Delta w_{ij} = \lambda \sigma_i \sigma_j \] Hebbian

\[ \Delta w_{ij} = \frac{1}{N} \left( \sigma_i \sigma_j - \frac{1}{N} \sigma_i h_{ij} - \frac{1}{N} h_{ij} \sigma_j \right) \]

with \( h_{ij} = \sum_{k=1}^{n} w_{ik} y_k \) Storkey
ONN On-Chip Learning Architecture

Zynq board:
Learning: ARM processor (PS)
Inference: Logical resources (PL)
Tests on a 25-neuron ONN
On-Chip Learning

Zynq board:
- Learning: ARM processor
- Inference: Logical resources
On-Chip Learning

Results:
Zybo-Z7 board
5x3 ONN for digits recognition
• 3 training patterns
• 15 test patterns

Training patterns

Test patterns associated to digit ‘0’

Output (LEDs)

On-Chip Learning Results:
Zybo-Z7 board
5x3 ONN for digits recognition
• 3 training patterns
• 15 test patterns

Input (switches)
Train (button)

Learning

ONN in
Weights
Synapses
ONN out

ZYNQ Processor

TU/e
On-Chip Learning

Accuracy:

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Hebbian</th>
<th>Storkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>On-chip</td>
<td>93,33%</td>
<td>93,33%</td>
</tr>
<tr>
<td>Off-chip</td>
<td>93,33%</td>
<td>93,33%</td>
</tr>
</tbody>
</table>

On-Chip Learning

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Hebbian</th>
<th>Storkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>On-chip</td>
<td>93.33%</td>
<td>93.33%</td>
</tr>
<tr>
<td>Off-chip 1</td>
<td>93.33%</td>
<td>93.33%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resources</th>
<th>Hebbian</th>
<th>Storkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>On-chip</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- LUTs</td>
<td>8203 (15.42%)</td>
<td>8203 (15.42%)</td>
</tr>
<tr>
<td>- Flip-Flops</td>
<td>3305 (3.11%)</td>
<td>3305 (3.11%)</td>
</tr>
<tr>
<td>Off-chip 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- LUTs</td>
<td>958 (1.8%)</td>
<td>800 (1.5%)</td>
</tr>
<tr>
<td>- Flip-Flops</td>
<td>721 (0.68%)</td>
<td>721 (0.68%)</td>
</tr>
</tbody>
</table>

On-Chip Learning

Results:

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Hebbian</th>
<th>Storkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>On-chip</td>
<td>93.33%</td>
<td>93.33%</td>
</tr>
<tr>
<td>Off-chip 1</td>
<td>93.33%</td>
<td>93.33%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resources</th>
<th>Hebbian</th>
<th>Storkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>On-chip</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- LUTs</td>
<td>8203 (15.42%)</td>
<td>8203 (15.42%)</td>
</tr>
<tr>
<td>- Flip-Flops</td>
<td>3305 (3.11%)</td>
<td>3305 (3.11%)</td>
</tr>
<tr>
<td>Off-chip 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- LUTs</td>
<td>958 (1.8%)</td>
<td>800 (1.5%)</td>
</tr>
<tr>
<td>- Flip-Flops</td>
<td>721 (0.68%)</td>
<td>721 (0.68%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Learning time</th>
<th>Hebbian</th>
<th>Storkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>On-chip</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Computation</td>
<td>33 us</td>
<td>77 us</td>
</tr>
<tr>
<td>- Transmission</td>
<td>86 us</td>
<td>86 us</td>
</tr>
</tbody>
</table>

On-Chip ONN Learning Results

Precision:
- 25-neuron ONN trained with random patterns and tested with random input images with various hamming distances
- Comparison with HNN on Matlab
- With HNN, Hebbian has lower capacity than Storkey, while with ONN we obtain similar results
- Reducing the weight precision does not impact much the precision.
On-Chip ONN Learning Results

Resources utilization:
• Adding on-chip learning to ONN increases importantly the resource utilization.
• Maximum ONN size with on-chip learning: 35 neurons (fully-connected).
• Reducing weight precision does not reduce the resource utilization meaningfully.
On-Chip ONN Learning Results

Latency:

- Inference latency is the same with/without on-chip learning
- Reducing weight precision and ONN size reduce the transmission latency during learning process.
- Storkey requires more computation and takes more time to compute
- Training can take 350µs and 550µs for 25-neuron ONN.

<table>
<thead>
<tr>
<th></th>
<th>This work</th>
<th>(Abernot et al. 2022b)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Weights</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hebbian learning</td>
<td>3 bits: 55 µs</td>
<td>33 µs</td>
</tr>
<tr>
<td></td>
<td>4 bits: 210 µs</td>
<td>77 µs</td>
</tr>
<tr>
<td></td>
<td>5 bits: 140 µs</td>
<td>NA</td>
</tr>
<tr>
<td>Storkey learning</td>
<td>18 µs</td>
<td>71 µs</td>
</tr>
<tr>
<td></td>
<td>213 µs</td>
<td>266 µs</td>
</tr>
<tr>
<td></td>
<td>368 µs</td>
<td>421 µs</td>
</tr>
<tr>
<td>Weight precision</td>
<td>18 µs</td>
<td>71 µs</td>
</tr>
<tr>
<td></td>
<td>213 µs</td>
<td>266 µs</td>
</tr>
<tr>
<td></td>
<td>368 µs</td>
<td>421 µs</td>
</tr>
<tr>
<td>Weight transmission</td>
<td>18 µs</td>
<td>71 µs</td>
</tr>
<tr>
<td></td>
<td>213 µs</td>
<td>266 µs</td>
</tr>
<tr>
<td></td>
<td>368 µs</td>
<td>421 µs</td>
</tr>
<tr>
<td><strong>Total Hebbian</strong></td>
<td>213 µs</td>
<td>266 µs</td>
</tr>
<tr>
<td></td>
<td>368 µs</td>
<td>421 µs</td>
</tr>
<tr>
<td><strong>Total Storkey</strong></td>
<td>368 µs</td>
<td>421 µs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>163 µs</td>
</tr>
<tr>
<td><strong>Inference</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Input transmission</td>
<td>9 µs</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>17 µs</td>
<td>NA</td>
</tr>
<tr>
<td>Output transmission</td>
<td>18 µs</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>44 µs</td>
<td>NA</td>
</tr>
</tbody>
</table>
OAV with ONN

- Obstacle avoidance with an Arduino robot

1st-level ONN to detect obstacles

2nd-level ONN to define direction: Turn left

Sensor values (Thermometer encoding)

ONN – 5x8

ONN – 1x8

Directions

Proximity sensors x8

$I^2C$ reading

Data conversion to 5x8 ONN Input

AXI

PL ONN 5x8 ONN 1x8

SoC Zynq Processor

Control board

Drives robot

$I^2C$ reading

ONN output to direction

Detected obstacles
1st-level ONN to detect obstacles

256 training patterns showing robot environment map

2nd-level ONN to define direction: Turn left

16 training patterns associated to 3 directions

Thermometer encoding

Sensor values

Direction

TU/e
OAV with ONN

- **Obstacle avoidance with an Arduino robot**

1st-level ONN to detect obstacles

2nd-level ONN to define direction: Turn left

- **ONN ON-CHIP LEARNING**
  - with a 15-neuron ONN on the Zynq processor
  - Hebbian learning rule
  - Up to 3 training patterns
On-Chip Learning for Obstacle Avoidance

- Development of a first ONN-based system integrating on-chip learning as a feedback loop for a real-world obstacle avoidance application.

- Two solution, with or without prior knowledge are validated to train ONN to avoid obstacles from environment.
OAV on ONN with On-Chip Learning

![Graph showing accuracy over time](image1)

### Accuracy (%)

- **Accuracy TP**
- **Accuracy direction**

#### Weight update

- 'right'
- 'front'
- 'left'

![Graph showing weight update](image2)

#### Input samples through time (t)

<table>
<thead>
<tr>
<th>ONN-level</th>
<th>ONN 1</th>
<th>ONN 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (# neurons)</td>
<td>40</td>
<td>8</td>
</tr>
<tr>
<td>Synapses resolution</td>
<td>5 signed bits</td>
<td>5 signed bits</td>
</tr>
<tr>
<td>LUTs (%)</td>
<td>4529 (25.7%)</td>
<td></td>
</tr>
<tr>
<td>Flip-Flops</td>
<td>3155 (9.1%)</td>
<td></td>
</tr>
<tr>
<td>Training</td>
<td>Pre-trained</td>
<td>Incremental</td>
</tr>
<tr>
<td>Accuracy</td>
<td>100 %</td>
<td>NA</td>
</tr>
</tbody>
</table>

#### System-level ($F_{PS} = 325 MHz$)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LUTs (%)</td>
<td>5183 (29.5%)</td>
</tr>
<tr>
<td>Flip-Flops (%)</td>
<td>4127 (11.7%)</td>
</tr>
<tr>
<td>Inference Latency</td>
<td></td>
</tr>
<tr>
<td>- Sensor measurement</td>
<td>18 ms</td>
</tr>
<tr>
<td>- Transmission</td>
<td>122 μs</td>
</tr>
<tr>
<td>- ONN</td>
<td>51 μs</td>
</tr>
<tr>
<td>Training latency</td>
<td></td>
</tr>
<tr>
<td>- Hebbian</td>
<td>13 μs</td>
</tr>
<tr>
<td>- Transmission</td>
<td>495 μs</td>
</tr>
<tr>
<td>- Post-processing</td>
<td>1.85 μs – 3.1 μs</td>
</tr>
</tbody>
</table>
On-Chip Learning **With** Prior Knowledge
On-Chip Learning **Without** Prior Knowledge
Conclusion

• First solution to provide on-chip learning to ONN implemented on FPGA compatible with various learning rules.

• Limitations
  • Scalability: the ONN size is limited by the resources utilization (LUTs) due to reconfigurable synapses to maximum 35 fully-connected neurons.

• Next steps:
  • Explore alternative solutions to improve scalability.
  • Explore possible applications: robotics.
Thank you!
Copyright Notice

This multimedia file is copyright © 2023 by tinyML Foundation. All rights reserved. It may not be duplicated or distributed in any form without prior written approval.

tinyML® is a registered trademark of the tinyML Foundation.

www.tinyml.org
Copyright Notice

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