On Device Learning Forum

Enabling Ultra-low Power Machine Learning at the Edge

“Merging insights from artificial and biological neural networks for neuromorphic edge intelligence”

Charlotte Frenkel – Assistant professor, Delft 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/
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
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 Dellafererra "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

Pacific Time
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

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
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.

Where what if becomes what is.
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

Test & Validation

Code Generation

Code Splitting

Version Control and Model Management

Team Collaboration

sensiml.com/tinyML
STMicroelectronics provides extensive solutions to make tiny Machine Learning easy
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:

tinyML - Enabling ultra-low Power ML at the Edge

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
On device learning Forum

tinyML EMEA Innovation Forum

June 26 -28, 2023
Amsterdam

EMEA 2023
https://www.tinymce.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
Charlotte Frenkel received the Ph.D. degree from Université catholique de Louvain (UCLouvain), Belgium, in 2020. After a postdoc at the Institute of Neuroinformatics, UZH and ETH Zürich, Switzerland, she joined Delft University of Technology, The Netherlands, as an Assistant Professor in July 2022. Her research focuses on neuromorphic integrated circuit design and learning algorithms for adaptive edge computing. She received a best paper award at the IEEE ISCAS 2020 conference, as well as the FNRS Nokia Bell Labs Scientific Award, the FNRS IBM Innovation Award and the UCLouvain/ICTEAM Best Thesis Award for her Ph.D. thesis. She serves as a TPC member for the tinyML Research Symposium and for the IEEE ESSCIRC, ISLPED, and DATE conferences.
Merging insights from artificial and biological neural networks for neuromorphic intelligence

Charlotte Frenkel

Delft University of Technology, Microelectronics Department
c.frenkel@tudelft.nl

tinyML On-Device Learning Forum 2023
Online, May 16th 2023
Outline

1. From neuroscience to AI and back again... ...which perspective?
   ...which starting point?

2. Why should we bother with neuroscience?

3. How can we morph these questions into interesting solutions for on-device-learning?
From neuroscience to AI and back again
Which starting point? Which perspective?

AI without hardware is unsustainable

Science
Neuroscience
Algorithms
Engineering
Neuromorphic hardware

Frenkel, tinyML ODL Forum 2023
Outline

1. From neuroscience to AI and back again... which perspective?
2. Why should we bother with neuroscience?
3. How can we morph these questions into interesting solutions for on-device-learning?
Synaptic plasticity rules – Neuroscience as the starting point

Synergy with hardware: the perspective of data locality

Spike-timing-dependent plasticity (STDP)

[Bi & Poo, J. Neurosci., 1998]
[Cassidy, ISCAS, 2011]

Spike-dependent synaptic plasticity (SDSP)

[Brader, Neur. Comp., 2007]

Local

Local
Synaptic plasticity rules – Neuroscience as the starting point

Synergy with hardware: the perspective of data locality

Frenkel, Trans. BioCAS, 2019

STDP

SDSP

[Cassidy, ISCAS’11]
Synaptic plasticity rules – Neuroscience as the starting point

Synergy with hardware: the perspective of data locality

Spike-timing-dependent plasticity (STDP)

\[ \Delta w \]

\[ \Delta (t_{\text{post}} - t_{\text{pre}}) \]

- Local in space
- Non-local in time

Spike-dependent synaptic plasticity (SDSP)

\[ \Delta w \]

- Local in space
- Local in time

Huge savings in silicon

\[ V_{\text{mem}} \geq \theta_m \]

\[ V_{\text{mem}} < \theta_m \]

[Brader, Neur. Comp., 2007]

[Bi & Poo, J. Neurosci., 1998]

[Cassidy, ISCAS, 2011]

Synaptic plasticity rules – Neuroscience as the starting point

Synergy with hardware: the perspective of data locality

ODIN (single-core)

- Record synaptic density
- Energy efficiency competitive with analog mixed-signal designs
- Large feature set (incl. synaptic plasticity)
  ...but quite painful to exploit!

MorphIC (quad-core)

AI algorithms

Neuroscience

Frenkel, tinyML ODL Forum 2023
Neural network training – Bio-plausibility as the end goal

Synergy with hardware: latency, memory access patterns

Output-independent target signals are also found in the brain!

Only ~15% overhead in power and area
[Frenkel, ISCAS'20] (🏆 Best paper award)
Designing efficient hardware hints toward bio-plausible mechanisms

Bringing AI closer to neuroscience leads to hardware efficiency

HW efficiency and bio-plausibility are often two sides of the same coin!
Many more examples: quantization, stochastic computing, event-driven computation,...
From neuroscience to AI and back again

Which starting point? Which perspective?

AI without hardware is unsustainable

1. Bottom-up science-driven approach
   - Analysis-by-synthesis
   - Difficult to scale efficiently to real-world problems

2. Top-down engineering-driven approach
   - Starts from working solutions to real-world problems
   - Which “salt & pepper” from neuroscience?

Neuromorphic intelligence:

2 should be fed by 1
Outline

1. From neuroscience to AI and back again... ...which perspective? ...which starting point?

2. Why should we bother with neuroscience?

3. How can we morph these questions into interesting solutions for on-device-learning?
Why is on-chip learning over second-long timescales difficult?

Let’s solve a yet unsolved engineering challenge!

Key challenge: End-to-end on-chip learning over long timescales while keeping a fine-grained temporal resolution

- Unrolling in time: very deep network (current learning ICs for static stimuli: ≤3 layers)
- Intractable memory/latency requirements
- No end-to-end on-chip solution to date (you still need costly external memory!)

Obviously, the brain doesn’t do that!
The bio-inspired solution

Backprop through time (BPTT, backward)

Eligibility propagation (e-prop, forward)

Backward vs. forward-mode training

Biological plausibility
Space and time locality
On-chip memory requirements
Algorithmic developments toward efficient long-term on-chip training

Network definitions and evaluation task

Network model

Delaye-d-supervision navigation task

Sample input

Leaky integrate-and-fire (LIF)

Leaky integrator (LI)

Time [s]

"Left" cues

"Right" cues

Action prompt

Noise

X

Z

Y

0

40

20

0

40

20

1

0

0.0

0.5

1.0

1.5

2.0

Time [s]

Cue tracking

Working memory

Delayed supervision

Left

Right

Action

Frenkel, tinyML ODL Forum 2023
Algorithmic developments – Step 1

Neuron model selection

Network model

Sample input

Leaky integrate-and-fire (LIF) or adaptive LIF (ALIF)

Leaky integrator (LI)

Task performance

Test-set decision error rate

BPTT - 20-ms-leak LIF
BPTT - Adaptive LIF

Frenkel, tinyML ODL Forum 2023
Algorithmic developments – Step 1

Neuron model selection

Network model

Task performance

Leaky integrate-and-fire (LIF):
- Only a short time constant (~20-ms leak)

Adaptive LIF (ALIF):
- Embeds threshold adaptation over 100s of ms

Frenkel, tinyML ODL Forum 2023
Algorithmic developments – Step 1

Neuron model selection

Leaky integrate-and-fire (LIF):
- Only a short time constant (~20-ms leak)
- Eligibility traces: simple activity LPF

Adaptive LIF (ALIF):
- Embeds threshold adaptation over 100s of ms
- ET: Complex per-synapse multi-scale filtering

LIF with configurable leak:
- Flexible time constant (ms to sec)
- Eligibility traces: simple activity LPF
- Less biologically plausible
Algorithmic developments – Step 2

Space and time locality

\[ \frac{dE}{dW_{ij}} \approx \sum_t L_j^t e_{ji}^t \]

Local decoupling of space and time

Step 3 – Stochastic weight updates allow reducing weight resolution to 8 bits.
- Competitive with a BPTT-trained network of ALIF neurons
- Memory overhead reduced to 0.8% of the inference-only network
From neuroscience to AI and back again
Which starting point? Which perspective?

Neuromorphic intelligence: ② should be fed by ①
ReckOn – Neuroscience and AI meet in efficient hardware

Chip microphotograph and summary

<table>
<thead>
<tr>
<th>Technology</th>
<th>28nm FDSOI CMOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core size</td>
<td>0.67 x 0.67 mm²</td>
</tr>
<tr>
<td>Die size</td>
<td>0.93 x 0.93 mm²</td>
</tr>
<tr>
<td>SRAM</td>
<td>138kB + 0kB ext. DRAM!</td>
</tr>
<tr>
<td>Network</td>
<td>Spiking RNN</td>
</tr>
<tr>
<td>Training timespan</td>
<td>Max. 32k steps</td>
</tr>
</tbody>
</table>

- Controller FSM, SPI and AE interfaces, neuron update logic, weight update logic
- W_{out} SRAM (8kB)
- Neur SRAM (2kB)
- W_{rec} SRAM (64kB)
- W_{inp} SRAM (64kB)

Frenkel, tinyML ODL Forum 2023

[Frenkel, ISSCC, 2022]
ReckOn – Neuroscience and AI meet in efficient hardware
Chip microphotograph and summary

- Event-driven / sparsity-aware computation
- Sensor-agnostic raw-data processing
- Task-agnostic processing and learning

Frenkel, ISSCC, 2022
ReckOn – Measurements and benchmarking
Three benchmarks that demonstrate task-agnostic learning

Vision
IBM DVS Gestures dataset
(10 classes, shrunk to 16x16)

Audition
Spiking Heidelberg Digits (EN) dataset
(1-word KWS, 1:1 target vs. filler, 1:3 sub)

Navigation
Delayed-supervision cue accumulation

Accuracy: 87.3% (28µW @0.5V)
Accuracy: 90.7% (46µW @0.5V)
Accuracy: 96.4% (14µW @0.5V)

[Frenkel, ISSCC, 2022]
ReckOn – What you should remember

Key elements toward neuromorphic edge intelligence

ReckOn merges AI, neuroscience and hardware to

achieve end-to-end on-chip learning over second-long timescales while keeping a milli-second temporal resolution, a yet unsolved challenge,

provide a low-cost solution: 0.45-mm² core area, <50µW for training at 0.5V, and 0.8-% memory overhead vs. inference-only network,

demonstrate task-agnostic learning with a spike-based encoding toward user customization and chip repurposing at the edge.

Neuromorphic intelligence outlines an exciting future for tiny on-device learning!
The Cognitive Sensor Nodes and Systems (CogSys) Team

We bridge the bottom-up (bio-inspired) and top-down (engineering-driven) design approaches toward neuromorphic intelligence.

Things we like (non-exhaustive list!):

• designing neuromorphic ICs and tinyML accelerators (mostly digital, going mixed-signal)
• bio-plausible training algorithms and synaptic plasticity mechanisms
• system-level optimization for autonomous sensorimotor agents, from sensing to decision

Positions will open soon!
Main references:


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