tinyML Neuromorphic Engineering Forum

https://www.tinyml.org/event/tinyml-neuromorphic-engineering-forum/

September 27, 2022

Introduction

tinyML is a fast-growing initiative around low-power machine-learning technologies for edge devices. The scope of tinyML naturally aligns with the field of neuromorphic engineering, whose purpose is to replicate and exploit the way biological systems sense and process information within constrained resources.

In order to build on these synergies, we are excited to announce the first tinyML Forum on Neuromorphic Engineering. During this event, key experts from academia and industry will introduce the main trends in neuromorphic hardware, algorithms, sensors, systems, and applications.

tinyML Neuromorphic Technical Program Committee

Charlotte Frenkel, Chair, Delft University of Technology
Christoph Posch, PROPHESEE
Jae-sun Seo, Arizona State University
Priya Panda, Yale University
Sadique Sheik, SynSense AG
Yulia Sandamirskaya, Intel
Friedemann Zenke, University of Basel
Andre van Schaik, Western Sydney University
Evgeni Gousev, Qualcomm
Ira Feldman, tinyML Foundation
Bette Cooper, tinyML Foundation
Olga Goremichina, tinyML Foundation

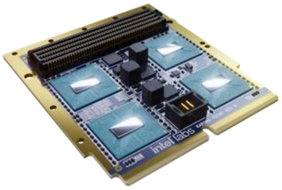
Note: The order of talk overviews is arranged according to the schedule of the event.

Neuromorphic Intelligence and Learning in Robotics

Yulia Sandamirskaya, Applications Research Lead, Neuromorphic computing lab, yulia.sandamirskaya@intel.com

Overview

Assistive human-centered robotics is a seminal use case for power-, time-, and data-efficient AI: autonomous mobile robots, robotic arms, and humanoids will have no time, space, or energy to lose when solving the recognition, localization, planning, decision making, and control tasks in dynamic and unstructured human-centered environments. Neuromorphic hardware and computing framework promise to support a variety of neural network-based “algorithms” to enable efficient AI for autonomous systems: pattern recognition and learning, parallel and graph-based search, model-predictive and adaptive control, optimization, planning, and skill learning. How do we identify and combine the most efficient neuromorphic algorithms? How can we program and evaluate them quickly and reliably? I will present a proposal for how neuromorphic algorithms can be developed and implemented. I will argue that the open-source software framework Lava can enable development of programs for neuromorphic hardware, exploiting the full potential of neuromorphic algorithms and leading to new, energy efficient and robust AI paradigm for autonomous assistive systems. I will show a couple recent examples of neuromorphic applications developed on Intel’s research chip Loihi.



Key highlights

- Neuromorphic hardware supports a wide range of neural network-based algorithms
- We need new algorithmic and programming paradigm to build neuromorphic applications
- Autonomous assistive robotics is one application that will profit from novel, neuromorphic AI

References and useful links

[1] https://github.com/lava-nc

[2] Hajizada, E., Berggold, P., Iacono, M., Glover, A., & Sandamirskaya, Y. (2022). Interactive continual learning for robots: a neuromorphic approach. In *Proceedings of the International Conference on Neuromorphic Systems 2022* (pp. 1-10) [https://dl.acm.org/doi/pdf/10.1145/3546790.3546791]

[3] Sandamirskaya, Y., Kaboli, M., Conradt, J., & Celikel, T. (2022). Neuromorphic computing hardware and neural architectures for robotics. *Science Robotics*, 7(67) [https://www.science.org/doi/abs/10.1126/scirobotics.abl8419]

[4] Davies, M., Wild, A., Orchard, G., Sandamirskaya, Y., Guerra, G. A. F., Joshi, P., ... & Risbud, S. R. (2021). Advancing neuromorphic computing with loihi: A survey of results and outlook. *Proceedings of the IEEE*, *109*(5), 911-934 [https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9395703]

[5] https://www.intel.com/content/www/us/en/newsroom/news/intel-unveils-neuromorphic-loihi-2-lava-software.html#gs.d5nwtw

The SpiNNaker neuromorphic computing platform

Steve Furber, ICL Professor of Computer Engineering, The University of Manchester, UK, steve.furber@manchester.ac.uk

Overview

SpiNNaker (a contraction of Spiking Neural Network Architecture) is a digital many-core neuromorphic computing platform designed primarily to support large-scale models of brain networks in biological real time. In conception for over 20 years and in construction for over 15 years, the million-core SpiNNaker machine at Manchester has been supporting an open neuromorphic computing service under the auspices of the EU Human Brain Project since April 2016, and has been used of real-time modelling of a detailed cortical microcircuit, cerebellar models, basal ganglia and other brain areas. In addition to its use in brain modelling, its real-time characteristics render it useful for neurorobotic and other engineering applications. A second generation SpiNNaker chip has been developed in collaboration with TU Dresden offering a 10x improvement in functional density and energy efficiency, and first silicon is currently being used to support software development. SpiNNaker2 offers state-of-the-art neuromorphic performance and efficiency in a very flexible configuration, building on a decade of experience of deploying SpiNNaker1.

Key highlights

- The world’s largest neuromorphic (brain-inspired) computing platform
- Openly available under the auspices of the EU Human Brain Project since April 2016
- Incorporates a million ARM (mobile phone) processors

References and useful links

[1] SpiNNaker: A Spiking Neural Network Architecture, now publishers (Open Access), http://dx.doi.org/10.1561/9781680836523
[2] http://apt.cs.manchester.ac.uk/projects/SpiNNaker/
[3] http://spinnakermanchester.github.io
[4] The SpiNNaker Project: https://ieeexplore.ieee.org/document/6750072
[5] https://iopscience.iop.org/article/10.1088/1741-2560/13/5/051001/meta

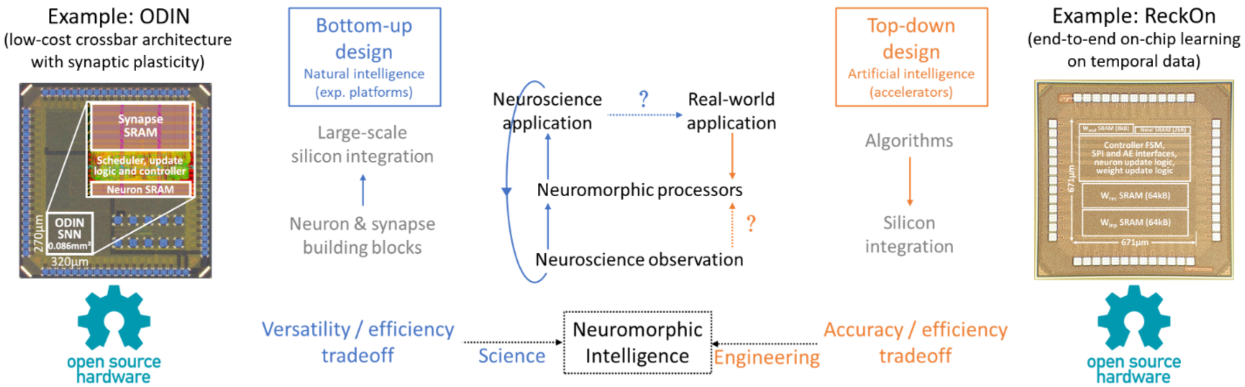
Digital Spiking Neural Network Accelerators for Neuromorphic Edge Intelligence

Charlotte Frenkel, Assistant Professor, Delft University of Technology, c.frenkel@tudelft.nl

Overview

A key question is often raised in neuromorphic chip design [1]: should we start from biological primitives and figure out how to apply them to real-world applications (bottom-up approach), or should we build on working AI solutions and modify them to increase their biological plausibility (top-down approach)?

I will briefly review the main use cases on each side and show that digital spiking neural network accelerators provide a flexible and efficient solution, which I will illustrate with two open-source digital neuromorphic chips: ODIN [2] and ReckOn [3]. Finally, I will highlight how neuromorphic intelligence can be a key enabler for edge computing applications.


Example: ODIN
(low-cost crossbar architecture with synaptic plasticity)
Synapse SRAM
Scheduler, update logic and controller
Neuron SRAM
ODIN SNN 0.086mm²
270µm
320µm
open source hardware
Bottom-up design
Natural intelligence (exp. platforms)
Large-scale silicon integration
Neuron & synapse building blocks
Versatility / efficiency tradeoff
Neuroscience application
?
Real-world application
Neuromorphic processors
?
Neuroscience observation
Science
Neuromorphic Intelligence
Engineering
Top-down design
Artificial intelligence (accelerators)
Algorithms
Silicon integration
Accuracy / efficiency tradeoff
Example: ReckOn
(end-to-end on-chip learning on temporal data)
Wout SRAM (8kB)
Neur SRAM (2kB)
Controller FSM, SPI and AE interfaces, neuron update logic, weight update logic
Wrec SRAM (64kB)
Winp SRAM (64kB)
671µm
671µm
open source hardware

Key highlights

- Digital neuromorphic chips offer an excellent tradeoff between robustness, flexibility, efficiency, scalability, and design time.
- A bottom-up approach best suits the design of neuromorphic experimentation platforms, while a top-down approach is preferred to demonstrate a competitive advantage in real-world applications.
- Top-down approaches need bottom-up insight toward neuromorphic intelligence.
- Digital processing is needed to support high-performance learning algorithms.

References and useful links

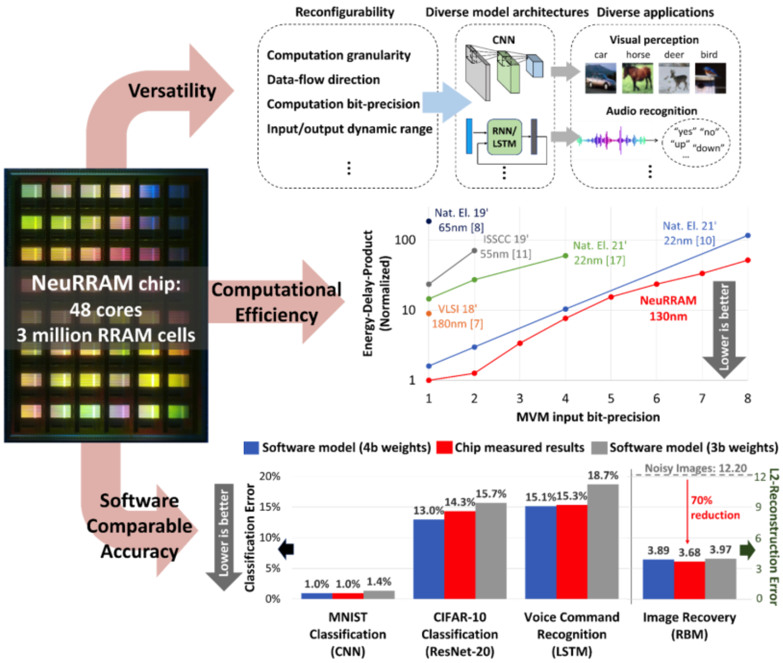
[1] C. Frenkel, D. Bol and G. Indiveri, *arXiv preprint arXiv:2106.01288,* 2021. https://doi.org/10.48550/arXiv.2106.01288
[2] C. Frenkel et al., *IEEE Trans. BioCAS*, vol. 13, no. 1, pp. 145-158, 2019. https://doi.org/10.1109/TBCAS.2018.2880425
Open-source repository: https://github.com/ChFrenkel/ODIN/
[3] C. Frenkel and G. Indiveri, *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2022. https://doi.org/10.1109/ISSCC42614.2022.9731734
Open-source repository: https://github.com/ChFrenkel/ReckOn/

Towards "Greener" AI on the Edge: Energy-Efficient Neuromorphic Learning and Inference

Gert Cauwenberghs, UC San Diego, gcauwenberghs@ucsd.edu

Overview

We present neuromorphic cognitive computing systems-on-chip implemented in custom silicon compute-in-memory neural and synaptic crossbar array architectures that combine the efficiency of local interconnects with flexibility and sparsity in global interconnects, and that realize a wide class of deeply layered and recurrent neural network topologies with embedded local plasticity for on-line learning, at a fraction of the computational and energy cost of implementation on CPU and GPGPU platforms. Co-optimization across the abstraction layers of hardware and algorithms leverage inherent stochasticity in the physics of synaptic memory devices and neural interface circuits with plasticity in reconfigurable massively parallel architecture towards high system-level accuracy, resilience, and efficiency. Adiabatic energy recycling in charge-mode crossbar arrays permit extreme scaling in energy efficiency, approaching that of synaptic transmission in the mammalian brain.



Key highlights

- **Superior energy efficiency** owing to extreme compute-in-memory parallelism, with neurosynaptic core crossbar elements operating near fundamental physical limits
- **High functional versatility** in configuring cores and their interconnectivity for diverse model architectures
- **High accuracy and resilience** through chip-in-the-loop training and fine-tuning, exploiting rather than fighting inherent device level nonlinearity and stochasticity

References and useful links

[1] G. Cauwenberghs, "Reverse Engineering the Cognitive Brain," *Proc. Nat. Acad. Sci. (PNAS)*, vol. **110** (39), pp. 15512-15513, 2013 (https://www.pnas.org/doi/10.1073/pnas.1313114110).

[2] W. Wan, R. Kubendran, C. Schaefer, S.B. Eryilmaz, W. Zhang, D. Wu, S. Deiss, P. Raina, H. Qian, B. Gao, S. Joshi, H. Wu, H.-S.P. Wong, and G. Cauwenberghs, "A Compute-in-Memory Chip Based on Resistive Random-Access Memory," *Nature,* vol. **608**, pp. 504–512, 2022 (https://www.nature.com/articles/s41586-022-04992-8).

[3] N. Mysore, G. Hota, S.R. Deiss, B.U. Pedroni, and G. Cauwenberghs, "Hierarchical Network Connectivity and Partitioning for Reconfigurable Large-Scale Neuromorphic Systems," *Frontiers in Neuroscience,* vol. **15**, pp. 797654:1-13, 2022 (https://www.frontiersin.org/articles/10.3389/fnins.2021.797654).

On-Sensor AI for Predictive Maintenance

Aleksandrs Timofejevs, CEO, atimofeev@polyn.ai

Overview

For Industry 4.0 health monitoring of machines is vital, and neural networks are ideal for that purpose. Vibration-based condition monitoring is one of the basic Predictive Maintenance options detecting machine failure.

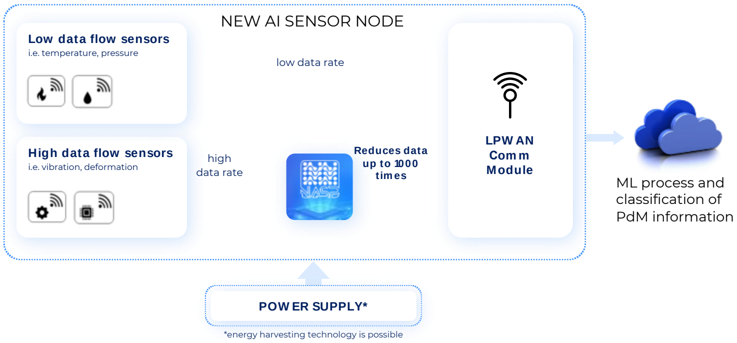
The power-hungry sensor node collects a lot of data for further analytics by Machine Learning (ML) algorithms. To send all this data to a center for analysis, the data communication would be more trouble than it worth. This shortens the battery life of operating sensor nodes. Data reduction can significantly decrease the volume of data sent to the cloud, saving OPEX and improving latency.

A NASP solution reduces that data flow by 1000 times, using neural network modeling, and enable to transmit through LoRa (or another wireless technology) only the embeddings extracted from the initial data. Thus, by applying a neural network, in this case on a NASP chip, we can obtain the whole range of patterns of vibration signals from various vibration sensors. The use of embeddings only reduce data sent to the cloud, solving the fundamental problem of low bandwidth required by IIoT systems.

NASP technology provides the optimal answer to power consumption and computing latency challenges through a hybrid analog and digital solution. It combines the advantages of the fixed weights part of the NASP chip and flexible weights in a digital co-processor, with smart optimization (pre-processing) of raw data directly on-sensor.

Key highlights

- TinyAI on sensor suitable for multi-year battery life or energy harvesting
- Full flexibility of AI model
- Data flow reduction 1000 times



References and useful links

[1] https://polyn.ai/technology-3/

[2] https://towardsdatascience.com/embeddings-beyond-just-words-2c835678dae2

[3] https://polyn.ai/polyn-technology-delivers-nasp-test-chip-for-tiny-ai/

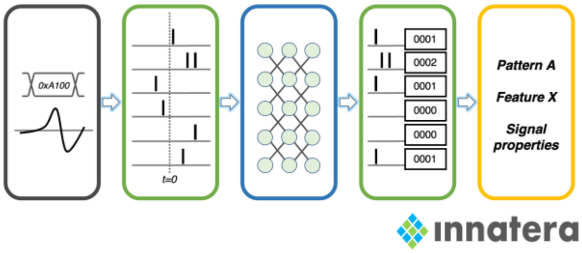
[4] https://polyn.ai/wp-content/uploads/2022/05/neuromorphic-analog-signal-processing-white-paper.pdf

Tiny spiking neural networks for sub-milliwatt AI at the sensor-edge

Sumeet Kumar, CEO, sumeet.kumar@innatera.com

Overview

Sensory data processing in the brain relies on event-driven networks of energy-efficient, continuous-time *analog* neurons and synapses. Spiking networks are able to perform advanced signal processing and AI functions even with tiny models due to their inherent notion of time. Innatera's Spiking Neural Processor (SNP) enables ultra-low power acceleration of these brain-inspired neural networks. The SNP uses an innovative processing architecture that implements spiking neural networks atop a continuous-time analog-mixed signal computational fabric, achieving unprecedented inference performance within a milliwatt-scale power envelope, and millisecond-scale latency. The SNP architecture is programmed like any conventional deep-learning accelerator, through Innatera's PyTorch-driven SDK named Talamo. The SDK radically simplifies the development of models through a turn-key workflow that eliminates the need for prior knowledge of spiking neural networks. In this talk, Innatera CEO Sumeet Kumar introduces the company's approach to sensor data processing with tiny spiking models, and how its Spiking Neural Processor enables ultra-low power inference at the edge. The talk briefly presents the Talamo SDK, and provides a comparison of neuromorphic versus standard deep-learning accelerators in the context of real-world applications.



Key highlights

- Future of tiny ML at the edge is neuromorphic – always-on processing with tiny spiking neural networks
- Ultra-low power inference on Innatera's Spiking Neural Processor
- Simplified, turn-key development of spiking neural networks with the Talamo SDK

References and useful links

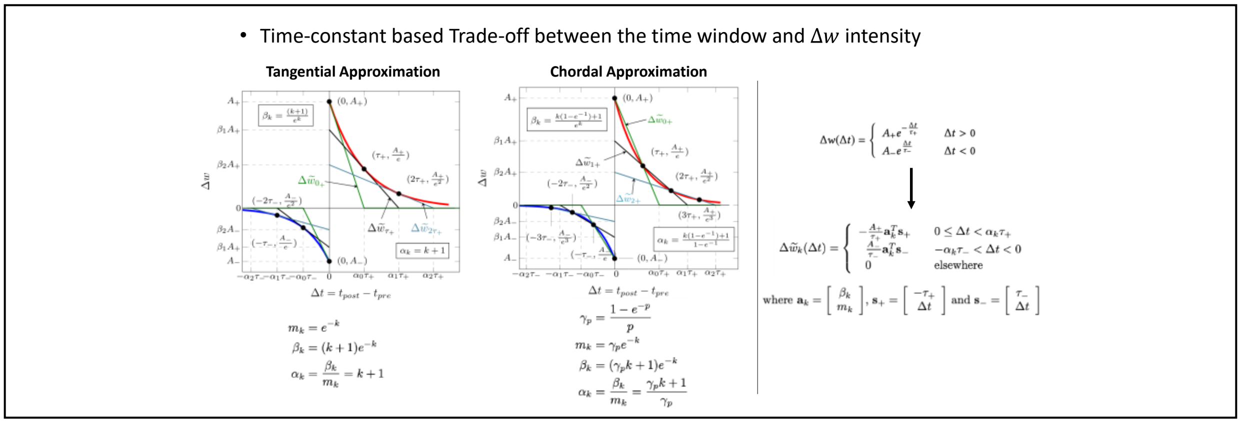
[1] www.innatera.com

Hardware Friendly Learning for Edge ML

Akwasi Akwaboah, Graduate Student, aakwabo1@jhu.edu; Ralph Etienne-Cummings, Professor, retienne@jhu.edu

Overview

Realizing Hebbian plasticity in large-scale neuromorphic systems is essential for reconfiguring synapses during recognition tasks. Spike-timing dependent plasticity (STDP), as a tool to this effect, has received a lot of attention in recent years. This phenomenon encodes weight update information as correlations between the presynaptic and postsynaptic event times, as such, it is imperative for each synapse in a silicon neural network to somehow understand track its activity and keep its own time. Carefully design synapses that can do that can be incorporated into compact, dense and energy efficient learning hardware. Here we present a biologically plausible and optimized Register Transfer Level (RTL), and algorithmic approach to realize Nearest-Neighbor STDP with temporal tracking and management handled by the postsynaptic dendrite on which the synapse sits. We adopt a time-constant based ramp approximation for ease of RTL implementation and incorporation in large-scale digital neuromorphic systems. We will describe the architecture, circuits, function of our hardware realizable STDP based learning system, and its application to neuroSLAM.



- Time-constant based Trade-off between the time window and Δw intensity

Tangential Approximation

Chordal Approximation

$$m_k = e^{-k}$$
$$\beta_k = (k+1)e^{-k}$$
$$\alpha_k = \frac{\beta_k}{m_k} = k+1$$

$$\gamma_p = \frac{1-e^{-p}}{p}$$
$$m_k = \gamma_p e^{-k}$$
$$\beta_k = (\gamma_p k+1)e^{-k}$$
$$\alpha_k = \frac{\beta_k}{m_k} = \frac{\gamma_p k+1}{\gamma_p}$$

$$\Delta w(\Delta t) = \begin{cases} A_+ e^{-\frac{\Delta t}{\tau_+}} & \Delta t > 0 \\ A_- e^{\frac{\Delta t}{\tau_-}} & \Delta t < 0 \end{cases}$$

$$\downarrow$$

$$\Delta\tilde{w}_k(\Delta t) = \begin{cases} -\frac{A_+}{\tau_+}\mathbf{a}_k^T\mathbf{s}_+ & 0 \leq \Delta t < \alpha_k\tau_+ \\ \frac{A_-}{\tau_-}\mathbf{a}_k^T\mathbf{s}_- & -\alpha_k\tau_- < \Delta t < 0 \\ 0 & \text{elsewhere} \end{cases}$$

where $\mathbf{a}_k = \begin{bmatrix} \beta_k \\ m_k \end{bmatrix}$, $\mathbf{s}_+ = \begin{bmatrix} -\tau_+ \\ \Delta t \end{bmatrix}$ and $\mathbf{s}_- = \begin{bmatrix} \tau_- \\ \Delta t \end{bmatrix}$

Key highlights

- STDP learning method for easy hardware Implementation
- Register Transfer Language description and available code at https://github.com/Adakwaboah/LODeNNS
- Applicable to SNN training and future implementation of neuroSLAM

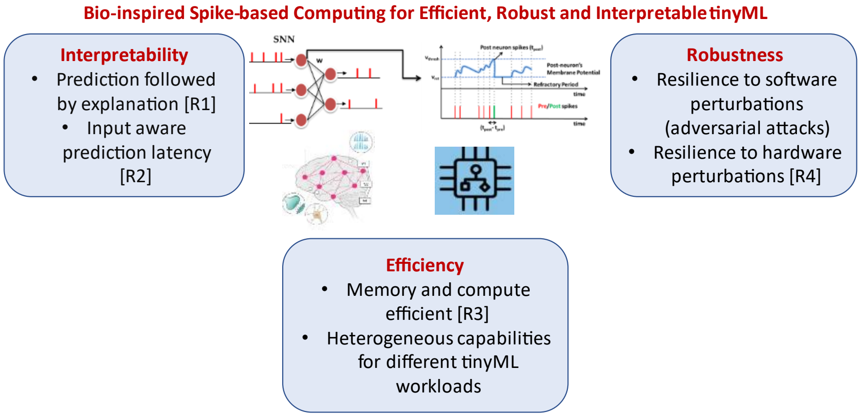
References and useful links

[1] A. D. Akwaboah and R. Etienne-Cummings, "LODeNNS: A Linearly-approximated and Optimized Dendrocentric Nearest Neighbor STDP," *ICONS '22: Proceedings of the International Conference on Neuromorphic Systems*, July 2022, Article No.: 3, Pages 1–8; https://doi.org/10.1145/3546790.3546793

Robust and Efficient tinyML with Spike-based Machine Intelligence

Priyadarshini Panda, Assistant Professor, Electrical Engineering, Yale University, USA
Email: priya.panda@yale.edu

Overview


Bio-inspired Spike-based Computing for Efficient, Robust and Interpretable tinyML

Interpretability
- Prediction followed by explanation [R1]
- Input aware prediction latency [R2]

Robustness
- Resilience to software perturbations (adversarial attacks)
- Resilience to hardware perturbations [R4]

Efficiency
- Memory and compute efficient [R3]
- Heterogeneous capabilities for different tinyML workloads

Spiking Neural Networks (SNNs) have recently emerged as an alternative to deep learning due to their huge energy efficiency benefits on neuromorphic hardware. However, training such models for diverse tasks for heterogeneous workloads central to tinyML tasks remains a challenge. In this presentation, I will talk about important techniques for training SNNs which bring a huge benefit in terms of latency, accuracy, interpretability, and robustness. I will first talk about **SNN training algorithms using neural architecture searches** at individual edge devices. Our recent work [1] finds SNNs search for brain-like feedback connections (interestingly, primate visual cortex has 20% feedforward and 80% feedback connections) that achieves state-of-the-art performance with significantly lesser latency and processing time. Similarly, using pruning can enable >90% weight sparsity that can be useful to deploy such models on extreme memory and power-constrained edge platforms. As SNNs find usage in real-world applications such as medical robots, and tiny drones, explainability in addition to performance is critical. I will talk about the **inter-spike interval-based visualization for SNNs** developed in our group [3] that does not require any compute expensive backpropagation to compute the attention map and therefore can be integrated on hardware at the edge. Essentially, this can potentially enable a paradigm of prediction followed by an explanation for all tinyML edge platforms. In the end, I will highlight the benefits of using SNNs for bringing in robustness such as **adversarial resilience, data/model privacy and the importance of building suitable hardware aware benchmarking platforms [3, 4]** (for von Neumann and emerging compute in memory architectures) to realize the energy and performance benefits of SNNs over traditional deep learning and their relevance for tinyML.

References & Links

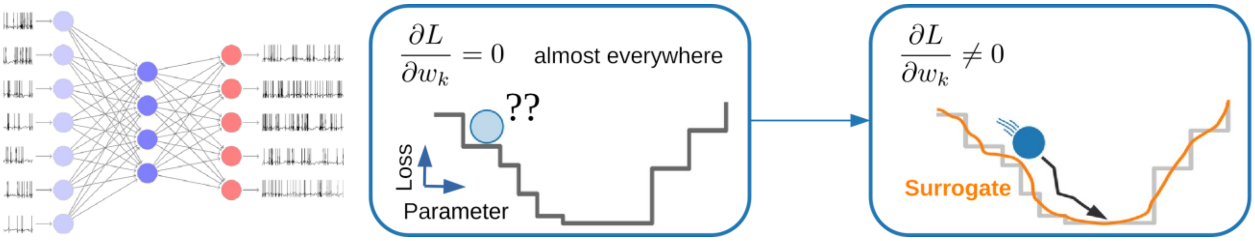
[R1] Kim et al. "Neural architecture search for spiking neural networks." *arXiv:2201.10355* (ECCV 2022). [R2] Kim et al. "Visual explanations from spiking neural networks using inter-spike intervals." *Scientific reports* 11.1 (2021): 1-14. [R3] Yin et al. "SATA: Sparsity-Aware Training Accelerator for Spiking Neural Networks." *arXiv:2204.05422* (IEEE TCAD 2022). [R4] Bhattacharjee, Abhiroop, et al. "Examining the Robustness of Spiking Neural Networks on Non-ideal Memristive Crossbars." ISLPED 2022 (Best Paper).
[L1] https://github.com/Intelligent-Computing-Lab-Yale
[L2] https://intelligentcomputinglab.yale.edu/publications

Training spiking neural networks end-to-end with surrogate gradients

Friedemann Zenke, Group Leader FMI, friedemann.zenke@fmi.ch

Overview

Brains rely on spiking neural networks for ultra-low-power information processing. Integrating similar efficiency into artificial intelligence requires learning algorithms to instantiate complex spiking neural networks and brain-inspired neuromorphic hardware to emulate them efficiently. To this end, I will briefly introduce surrogate gradients as a general framework for training spiking neural networks end-to-end [1], showcase its capabilities for instantiating spiking neural networks with sparse activity, and demonstrate its capabilities on analog neuromorphic hardware [2]. I will also outline a deep link between approximate surrogate gradients and a family of bio-inspired online learning rules [3-5].


$\frac{\partial L}{\partial w_k} = 0$ almost everywhere
??
Loss
Parameter
$\frac{\partial L}{\partial w_k} \neq 0$
Surrogate

Key highlights

- Reliable learning algorithms for training spiking neural networks end-to-end exist
- Surrogate gradients can self-calibrate analog neuromorphic substrates
- Effective online learning rules and initialization strategies inspired by neurobiology exist

References and useful links

[1] Neftci, E.O., Mostafa, H., and Zenke, F. (2019). IEEE SPM *36*, 51–63. https://doi.org/10.1109/MSP.2019.2931595.
[2] Cramer, B., Billaudelle, S., Kanya, S., Leibfried, A., Grübl, A., Karasenko, V., Pehle, C., Schreiber, K., Stradmann, Y., Weis, J., Schemmel, J., and Zenke F. (2022). PNAS *119*. https://doi.org/10.1073/pnas.2109194119.
[3] Rossbroich, J., Gygax, J., and Zenke, F. (2022). Fluctuation-driven initialization for spiking neural network training. https://doi.org/10.48550/arXiv.2206.10226.
[4] Zenke, F., and Ganguli, S. (2018). Neural Comput *30*, 1514–1541. https://doi.org/10.1162/neco_a_01086.
[5] Zenke, F., and Neftci, E.O. (2021). Proceedings of the IEEE *109*, 935–950. https://doi.org/10.1109/JPROC.2020.3045625.

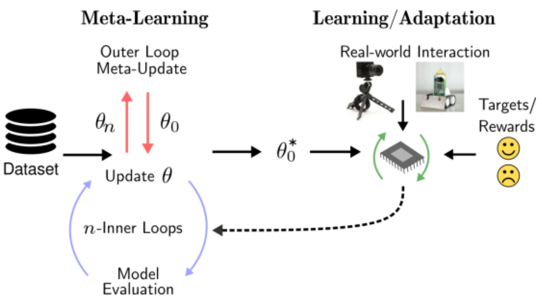
Enabling Neuromorphic Learning Machines with Meta-learning

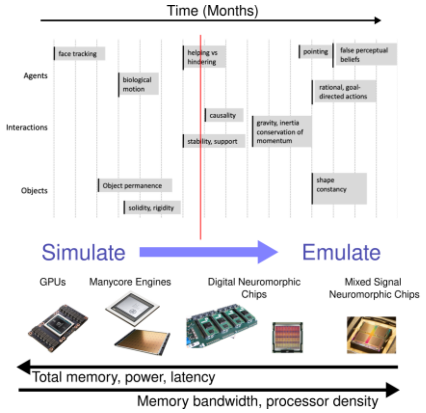
Emre Neftci, Forschungszentrum Jülich and RWTH Aachen

Overview

The data-intensive and randomized learning process that characterizes state-of-the-art Spiking Neural Network (SNN) training is incompatible with the physical nature and real-time operation of the brain and neuromorphic hardware. Bi-level learning, such as meta-learning, can be used in deep learning to overcome these limitations.

In this talk, I will focus on gradient-based meta-learning methods, namely Model Agnostic Meta Learning (MAML), in SNNs in conjunction with the surrogate gradient method and a roadmap of their implementation in neuromorphic hardware. I'll further discuss 1) the hardware advantages that accrue from meta-learning: fast learning without the requirement of high precision weights or gradients and training-to-learn with quantization and mitigating the effects of approximate synaptic plasticity rules, 2) the requirements with respect to datasets, and 3) and how meta-learning can enable new neuromorphic learning technologies for real-world problems.





Key highlights

- Real-world learning with low hardware requirements (precision, endurance, linearity, etc)
- Showcases applications for on-chip learning
- Requirements for dataset collection

References and useful links

[1] Stewart, Kenneth Michael, and Emre Neftci. "Meta-learning spiking neural networks with surrogate gradient descent." *Neuromorphic Computing and Engineering* (2022). (https://iopscience.iop.org/article/10.1088/2634-4386/ac8828/meta)

[2] Stewart, Kenneth, et al. "Online few-shot gesture learning on a neuromorphic processor." *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 10.4 (2020): 512-521.
(https://ieeexplore.ieee.org/abstract/document/9229141)

Neuromorphic Event-Based Vision

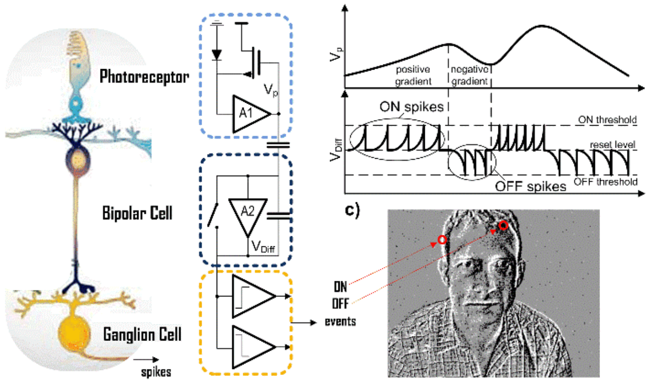
Christoph Posch, CTO, cposch@prophesee.ai

Overview

Neuromorphic Event-based (EB) vision is an emerging paradigm of acquisition and processing of visual information that takes inspiration from biology, trying to recreate its visual information acquisition and processing operations on VLSI silicon chips.

Today, the majority of EB sensor devices are based on the “temporal contrast” or “change detection” (CD) type of operation, loosely mimicking the transient Magno-cellular pathway of the human visual system. In contrast to conventional image sensors, CD sensors do not use one common sampling rate (=frame rate) for all pixels, but each pixel defines the timing of its own sampling points in response to its visual input by reacting to changes of the amount of incident light. The output generated by such a sensor is not a sequence of images but a quasi-time-continuous stream of pixel-individual contrast events, generated and transmitted conditionally, based on the dynamics happening in the scene. Acquired Information is encoded and transmitted in the form of data packets containing the originating pixel’s X,Y coordinate, time stamp, and often contrast polarity.

The highly efficient way of acquiring sparse data, the high temporal resolution and the robustness to uncontrolled lighting conditions are characteristics of the event sensing process that make EB vision attractive for numerous applications in industrial, surveillance, IoT, AR/VR, and automotive.



Key highlights

- High-speed vision at uncontrolled lighting conditions
- Sparse redundancy-free sampling
- Wide dynamic range operation

References and useful links

[1] P. Lichtsteiner, C. Posch and T. Delbruck, "A 128× 128 120 dB 15 μs Latency Asynchronous Temporal Contrast Vision Sensor," in IEEE Journal of Solid-State Circuits, vol. 43, no. 2, pp. 566-576, Feb. 2008

[2] C. Posch, T. Serrano-Gotarredona, B. Linares-Barranco and T. Delbruck, "Retinomorphic Event-Based Vision Sensors: Bioinspired Cameras With Spiking Output," in Proceedings of the IEEE, vol. 102, pp. 1470-1484, Oct. 2014

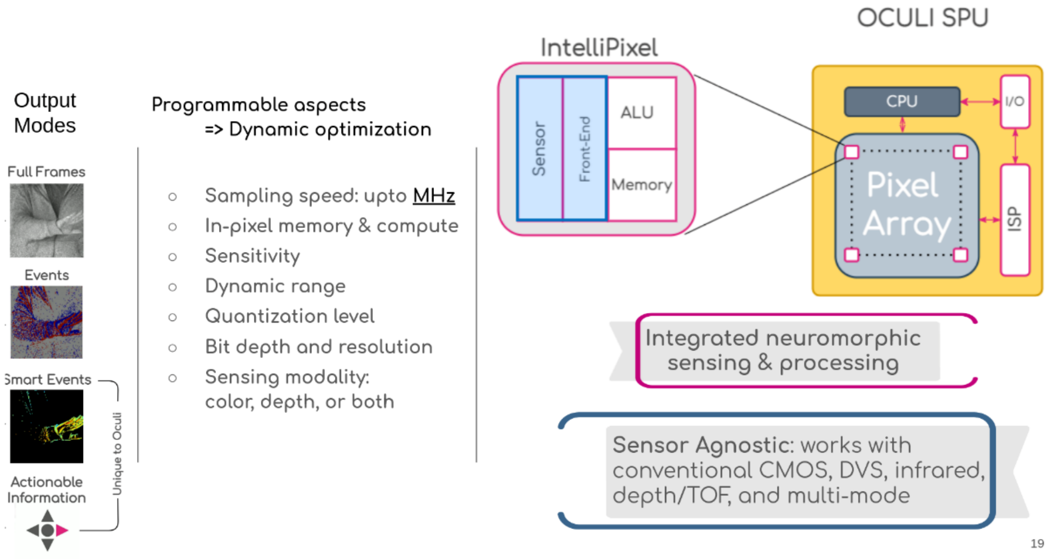
[3] T. Finateu et al., "5.10 A 1280×720 Back-Illuminated Stacked Temporal Contrast Event-Based Vision Sensor with 4.86μm Pixels, 1.066GEPS Readout, Programmable Event-Rate Controller and Compressive Data-Formatting Pipeline," ISSCC 2020

New Eyes Optimized For Machines

Dr. Charbel Rizk, CEO & Founder, charbel.rizk@oculi.ai

Overview

Oculi is putting the "Human Eye" in AI: machines outperform humans in most tasks but human vision remains far superior delivering the actionable signal in real time and consuming only mW's. As biology and nature have been the inspiration for much of the technology innovations, developing vision technology that mimics the eye+brain architecture is the logical path. Unlike photos and videos we collect for personal consumption, machine vision is not about pretty images and the most number of pixels. Machine vision should extract the “best” actionable information very efficiently (in time and energy) from the available signal (photons). At Oculi, we have developed a new architecture for computer and machine vision that enables dynamic and real time optimization. The core of this disruptive approach is the Oculi SPU (Sensing & Processing Unit) which is an intelligent Software Defined Vision Sensor combining sensing + processing at the pixel, the true edge for imaging sensors. This presentation will highlight the novel architecture and provide example use cases that are uniquely positioned for TinyML.



Programmable aspects
=> Dynamic optimization

- Sampling speed: upto MHz
- In-pixel memory & compute
- Sensitivity
- Dynamic range
- Quantization level
- Bit depth and resolution
- Sensing modality: color, depth, or both

Integrated neuromorphic sensing & processing

Sensor Agnostic: works with conventional CMOS, DVS, infrared, depth/TOF, and multi-mode

Key highlights

- Overview of the Oculi vision architecture.
- Vision Intelligence platform with S11 prototype.
- Example use cases for TinyML including the weather station challenge.

References and useful links

[1] https://globalventuring.com/university/oculi-eyes-computer-vision-revolution/
[2] https://www.edge-ai-vision.com/2022/05/edge-ai-and-vision-alliance-announces-2022-vision-tank-winners/
[3] https://www.eetimes.com/startup-mimics-human-eye-by-adding-processing-to-pixels/
[4] https://semiwiki.com/ceo-interviews/304902-ceo-interview-charbel-rizk-of-oculi/
[5] https://lebnet.us/Expert-Voice/11907027

Spiking Neural Networks for Low-Power Real-Time Inference

Sadique Sheik, VP, Artificial Intelligence, SynSense AG,
sadique.sheik@synsense.ai

Overview

SynSense is a low-power edge computing ASIC design company. We are developing approaches for machine vision, bio- and industrial signal processing based on neuromorphic principles. We specialize in integrated sensor-processor solutions for edge inference. As part of this endeavor, we also develop novel algorithms and solutions to find optimal spiking neural networks to solve real-world tasks.

This presentation will give you a brief overview of two of SynSense's device platforms a) Speck platform tailored for low-latency and ultra-low-power vision processing, b) Xylo platform tailored towards sub-mW low dimensional signal processing. We will have a look at the typical workflow from model development to deployment onto these devices and the corresponding python-based software infrastructure [2]. We will have a look at some commercial use cases where such technology could be applied with some demonstrators.

Key highlights

- Introduction to Integrated Sensor processors: Speck and Xylo
- Overview of the development pipeline.
- Use case demonstrations.

References and useful links

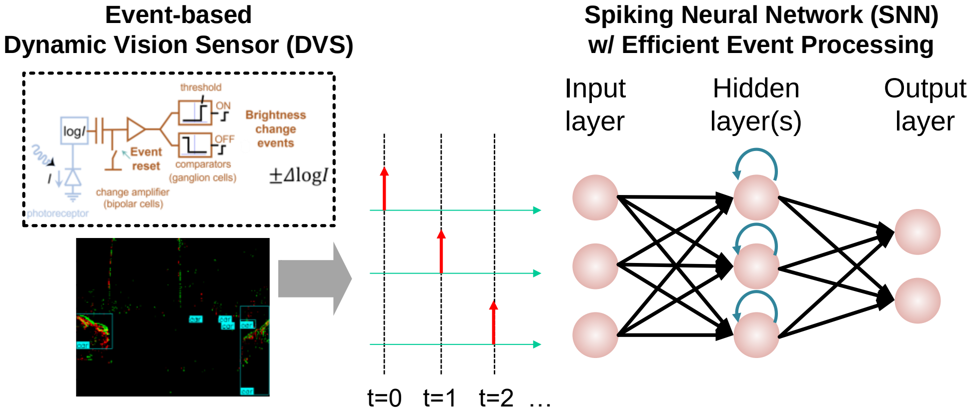
[1] synsense.ai
[2] sinabs.ai: Development pipeline for Speck, rockpool.ai: Development pipeline for Xylo

Fully Spike-based Processing with Front-end Dynamic Vision Sensor and Back-end Spiking Neural Network

Jae-sun Seo, Associate Professor, Arizona State Univ., jseo28@asu.edu

Overview

Spiking neural networks (SNN) mimic the operations in biological nervous systems. By exploiting event-driven computation and data communication, SNNs can achieve very low power consumption. However, two issues have persisted: (1) directly training SNNs have not resulted in competitive inference accuracy; (2) non-spike inputs (e.g. natural images) need to be converted to a train of spikes, which results in long latency. To exploit event-driven end-to-end operations, integration of spike-based front-end sensors such as dynamic vision sensors (DVS) and back-end SNNs become ideal. In addition, it is crucial to have a back-propagation based training algorithm that can directly train SNNs with continuous input spikes from DVS output. Such fully spike-based algorithm and hardware co-design will enable sparse and energy-efficient event-based end-to-end neuromorphic systems.


Event-based Dynamic Vision Sensor (DVS)

Spiking Neural Network (SNN) w/ Efficient Event Processing

Key highlights

- Combining front-end spike-based sensor such as dynamic vision sensor (DVS) and back-end spiking neural network (SNN) based accelerator fully exploits the end-to-end efficient event-based processing.
- SNN algorithm accuracy in recent literature has largely improved for image/video applications.
- Efficient SNN accelerators that tightly integrates with spike-based sensors are required.

References and useful links

[1] A. Lele et al., "An End-to-End Spiking Neural Network Platform for Edge Robotics: From Event-Cameras to Central Pattern Generation," IEEE Transactions on Cognitive and Developmental Systems, vol. 14, no. 3, Sep. 2022.
[2] J. Meng et al., “LT-SNN: Self-adaptive Spiking Neural Network with Learnable Threshold,” SRC TECHCON, 2022.

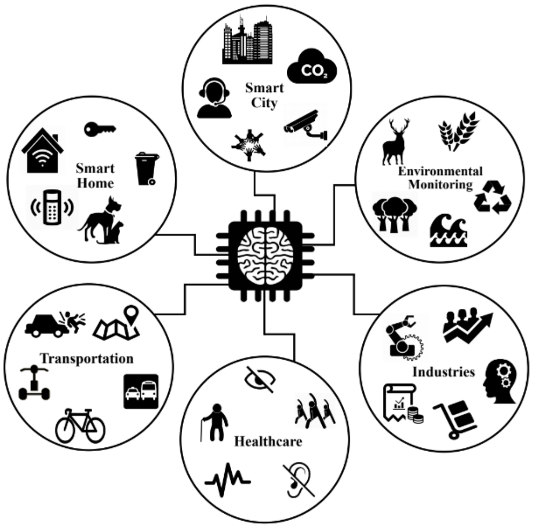
Indian Institute of Science
भारतीय विज्ञान संस्थान

tinyML In-filter Computing using Neuromorphic Cochlea

Chetan Singh Thakur, Assistant Professor, csthakur@iisc.ac.in

Overview

Edge devices are often constrained by the available computational power and hardware resource. We present a novel in-filter computing framework that can be used for designing ultra-light classifiers for time-series data. Unlike a conventional pattern recognizer, where the feature extraction and classification are designed independently, this architecture directly integrates convolution and nonlinear filtering operations into the kernels of a Support Vector Machine (SVM). The result of this integration is a template-based SVM with user-defined memory constraints in terms of fixed template vectors. Here, we have used the Neuromorphic Cochlea as a kernel in our template-based SVM formulation, which also acts as a feature extractor for time-series data. We prototyped the proposed system, on an FPGA and a Cortex-M4 MCU, for multiple ecological and healthcare applications using acoustic and IMU sensors.



Key highlights

- Neuromorphic Cochlea as a feature extractor and SVM kernel
- Sensing and inference at the edge within a few mW of power
- Portables to tiny FPGA and tiny microcontroller Boards
- Deployed as a sensor network for Ecological applications with LoRA connectivity

References and useful links

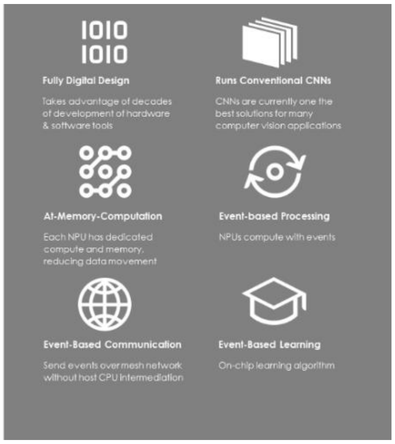
[1] A.R. Nair, S. Chakrabartty, C.S. Thakur, “*In-filter Computing For Designing Ultra-light Acoustic Pattern Recognizers*“, IEEE Internet of Things (IoT) Journal.
[2] H.R. Sabbella, A.R. Nair, V. Gumme, S.S. Yadav, S. Chakrabartty, C.S. Thakur, “*An Always-On tinyML Acoustic Classifier for Ecological Applications*,“ IEEE International Symposium on Circuits and Systems (ISCAS), 2022.
[3] tinyML Asia 2021 Video Poster: Bird Hotspots: A tinyML acoustic classification system for ecological insights.

Combining Neuromorphic Design Principles with Modern Machine Learning Algorithms

Anil Mankar, Chief Development Officer, amankar@brainchip.com

Overview

We discuss neuromorphic computing from the perspective of a company that designs computing solutions for machine learning (ML) applications at the edge. First, we examine how bringing ML applications to the edge is heavily influenced by the industry's past, which includes training deep neural networks (DNNs) on cloud servers, and how neuromorphic computing principles provide an excellent blueprint for solving constraints unique to edge computing. We then discuss how our design architecture applies neuromorphic principles to reduce the memory bandwidth and power usage at multiple scales while at the same time preserving important advantages that come with conventional ML algorithms and digital design. Finally, we discuss why neuromorphic computing is the future of edge computing and detail application areas and research topics where there is likely an opportunity for breakthroughs in efficiency and performance.



Fully Digital Design

Takes advantage of decades of development of hardware & software tools

Runs Conventional CNNs

CNNs are currently one the best solutions for many computer vision applications

At-Memory-Computation

Each NPU has dedicated compute and memory, reducing data movement

Event-based Processing

NPUs compute with events

Event-Based Communication

Send events over mesh network without host CPU intermediation

Event-Based Learning

On-chip learning algorithm

Key highlights

- It is possible to build fully digitally designed hardware that runs conventional ML algorithms using neuromorphic principles to dramatically reduce system-level memory bandwidth and power usage.
- Event-based computation and at-memory-compute are neuromorphic computing principles that can be applied in current hardware accelerator designs.
- CNNs provide state-of-the-art performance on many ML tasks but can be further augmented with neuromorphic inspired learning rules to enable on-chip learning and preserve customer privacy.

References and useful links

[1] https://brainchip.com/
[2] https://doc.brainchipinc.com/

WESTERN SYDNEY UNIVERSITY

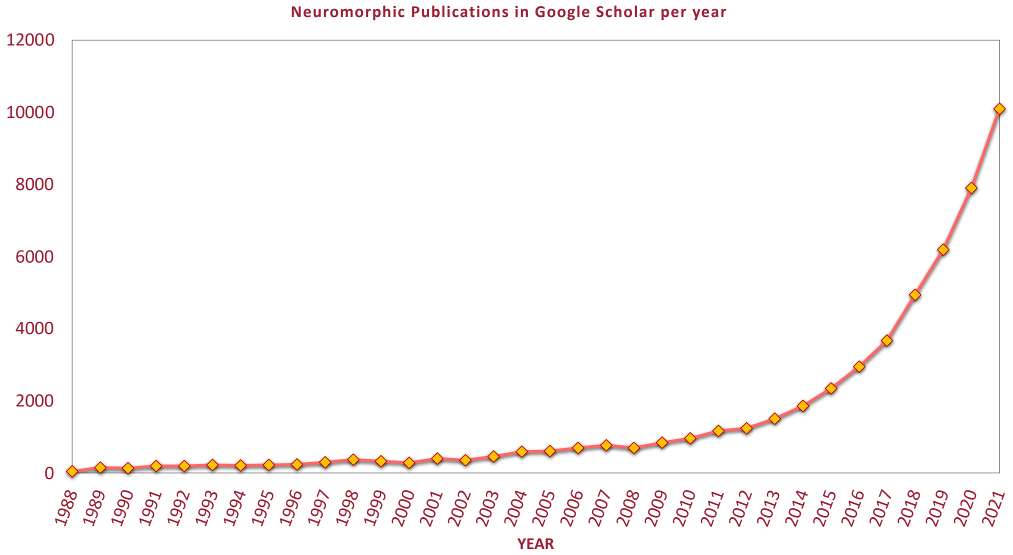
International Centre for Neuromorphic Systems

Neuromorphic Engineering needs applications

André van Schaik, Director International Centre for Neuromorphic Systems, a.vanschaik@westernsydney.edu.au

Overview

Interest in Neuromorphic Engineering has been growing exponentially, particularly in the past decade. I believe this is mainly due to the end of Moore's law, which has fuelled progress in electronics for over fifty years. Now that it is getting more and more difficult and expensive to cram ever more transistors on a single chip, the drive for industry to look for alternative approaches is strong, which in turn drives growth in academic research. Unlike Quantum computing, neuromorphic engineering promises to use currently available microelectronic manufacturing and design technology, but use it differently, rather than having to invent whole new manufacturing processes. The rise of smart devices, with strict size and power constraints, is also adding to the growth of interest in Neuromorphic Engineering. The 'smarts' in such devices are often that these devices don't just measure their environment, or operate independently of it, but instead perceive their environment and make decisions on how to act based on these observations. This is precisely what neural systems in biology have evolved for, hence the promise of neuromorphic engineering which takes its inspiration from such biological neural systems.



This growth in interest and funding is great for the field, but it is now up to us, Neuromorphic Engineers, to deliver on some of these promises and develop practical applications and do so within the next five years. Without such applications, I fear industry will conclude that Neuromorphic Engineering cannot deliver on its promises, and research funding will dry up. Thus, I argue for a strong applied research focus for the field, at least until we have demonstrated that we can indeed provide solutions to existing problems.

In this talk I will present some of the applications of neuromorphic technology we are developing at the International Centre for Neuromorphic Engineering and will highlight our world-first Master of Neuromorphic Engineering degree to train the next generation of Neuromorphic Engineers.

www.tinyML.org

The company and product information contained herein were provided by the respective companies. tinyML Foundation makes no assertion or guarantee as to the accuracy of this information. Inclusion in this market summary is not an endorsement by the tinyML Foundation or its sponsors of a specific company or technology. For participation in a future market summary please contact the tinyML Foundation.