Synaptic Normalisation for On-Chip Learning in Analog CMOS Spiking Neural Networks

Mastella, Michele; Greatorex, Hugh; Cotteret, Madison; Janotte, Ella; Soares Girão, Willian; Richter, Ole; Chicca, E.

Published in:
ACM ICONS2023

DOI:
10.1145/3589737.3606007

Publication date:
2023

Synaptic Normalisation for On-Chip Learning in Analog CMOS Spiking Neural Networks

Michele Mastella
Hugh Greatorex
m.mastella@rug.nl
University of Groningen; Cognigron
Groningen, The Netherlands

Madison Cotteret
Technische Universität Ilmenau
Ilmenau, Germany
University of Groningen; Cognigron
Groningen, The Netherlands

Ella Janotte
Istituto Italiano di Tecnologia
Genova, Italy
University of Groningen; Cognigron
Groningen, The Netherlands

Willian Soares Girão
Ole Richter
Elisabetta Chicca
University of Groningen; Cognigron
Groningen, The Netherlands

ABSTRACT

Spiking Neural Networks (SNNs) are becoming increasingly popular for their application in Edge Artificial Intelligence (Edge-AI) due to their sparse and low-latency computation. Among these networks, analog hardware SNNs are chosen for their ability to emulate complex dynamics in neurons and synapses, especially in integrated Metal Oxide Semiconductor (MOS) technology. They can form memories of external stimuli by modulating the strength of synaptic weights. In this context, binary weights are a common hardware design choice, due to their ease to program and store. The use of binary weights in SNNs worsens the bias introduced by the coding level of input stimuli (i.e. fraction of active input nodes), where the network activity is highly correlated to the number of excited neurons. In this paper, we present a Complementary Metal Oxide Semiconductor (CMOS) solution for the coding level bias, by proposing a novel circuit that employs synaptic normalisation at the neuron level. This circuit modifies the gain of the neuron depending on its input weights, with a small footprint and therefore high scalability.

CCS CONCEPTS

• **Hardware → Analog and mixed-signal circuits**; Integrated circuits; • **Computing methodologies → Neural networks**; **Bio-inspired approaches**.

Author Contribution: Conceptualisation - M.M.; Methodology - E.C., M.M.; Software/Hardware - E.J., H.G., M.C., M.M., O.R., W.S.G.; Investigation - H.G., M.M.; Writing - original draft - H.G., M.C., M.M; Writing – review and editing - E.C., E.J., O.R., W.S.G; Visualisation - H.G., M.C., M.M., O.R.; Supervision - E.C. (CRediT, authors listed in alphabetical order) .

ICONS '23, August 1–3, 2023, Santa Fe, NM, USA

ACM ISBN 979-8-4007-0175-7/23/08.
https://doi.org/10.1145/3589737.3606007

KEYWORDS

Synaptic Normalisation, Spiking Neural Networks, CMOS, Neuromorphic Engineering

ACM Reference Format:
Michele Mastella, Hugh Greatorex, Madison Cotteret, Ella Janotte, Willian Soares Girão, Ole Richter, and Elisabetta Chicca. 2023. Synaptic Normalisation for On-Chip Learning in Analog CMOS Spiking Neural Networks. In *International Conference on Neuromorphic Systems (ICONS '23), August 1–3, 2023, Santa Fe, NM, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589737.3606007

1 INTRODUCTION

The use of Application Specific Integrated Circuits (ASICs) for realising edge Edge-AI solutions is gaining popularity in the Internet of Things (IoT) community [5]. Among these solutions we can find analog hardware Spiking Neural Networks (SNNs), where neurons and synapses are designed to use the dynamics of transistors for replicating complex behaviours observed in biology. Due to the difficulties with implementing reliable programming for analog memories for these networks, weights are often constrained to binary values [7]. The coding level of input patterns, defined as the fraction of active input nodes, correlates heavily with the overall

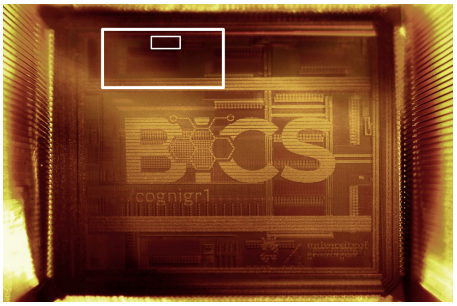

Figure 1: Photograph of the realised die. The small white rectangle indicates the area covered by the synaptic normalisation block, while the larger rectangle encompasses the peripheral supporting circuitry too.

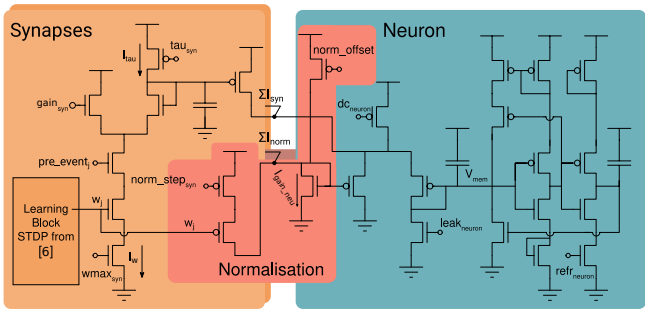

Figure 2: Schematic of the realised circuit. The ASIC implementation is composed of 2 synapses with learning capabilities and 1 neuron, all equipped with a synaptic normalization circuit. All capacitors are realised as MOS Capacitors (MOSCAPs).

network activity when using binary weights [2]. This results in unbalanced responses for different patterns, which is detrimental to learning and inference regimes. In fact, in the case where inference relies on the most active output neuron identification, the neuron tuned to the input stimuli with the largest coding level always dominates. Furthermore, the neurons tuned to input stimuli with lower coding level are constrained to exhibit a low activation. This can also bias the convergence of a learning algorithm towards input patterns with higher coding levels: the higher the coding scheme of an input, the higher the number of synapses encoding that pattern and therefore the chance of the pattern to be learnt, biasing the learning algorithm [2].

In this paper, we present a CMOS (Fig 1) implementation of synaptic normalisation, a method used to mitigate coding level effects by modulating the current flowing to the neuron depending on the number of active synapses.

2 METHODS

2.1 Circuit Description

As shown in Fig. 2, the realised circuit is composed of three distinct parts: one neuron and two synapses equipped with learning circuits. These components are inspired by circuits previously explored in literature [3]. They are, however, modified to incorporate the synaptic normalisation addition proposed in this work. In this section, the different circuit elements are briefly introduced in order to highlight how the synaptic normalisation circuit interacts with them. Note that the different voltage biases in the Fig. 2 are set by current mirrors. Therefore, in the following we will refer to the currents generating these voltages, rather than the voltages themselves (e.g. I_{w} instead of V_{w}). The only exception to this is $I_{\mathrm{gain_{syn}}}$, which is the virtual p-type subthreshold current biased by $V_{\mathrm{gain_{syn}}}$ [1].

2.1.1 DPI Synapse. The Differential Pair Integrator (DPI) circuit for implementing the synapses [1] exploits the trans-linear principle to obtain linear current behaviour with transistors operating in the sub-threshold regime. When a spike (i.e., a voltage pulse) arrives at the input (labelled as pre_event$_{\mathrm{j}}$ in Fig. 2) the current, set by I_w and gated by the digital signal w, flows through the differential pair and charges the capacitor (here realised as a n-type MOSCAP), resulting in an output current linearly dependent on the input spike frequency.

The current integrated by the capacitor depends on several biases: I_{w}, that limits the current present at the input branch, $I_{\mathrm{gain_{syn}}}$, that linearly increases the integrated current, I_{tau}, that defines a negative component of the integrated current which dominates in absence of an input pulse, and lastly w, that switches the synapse on or off. The average current sourced in response to an input spike train is calculated in [1] as:

$$\langle I_{\mathrm{syn}} \rangle = w \left(\frac{I_{\mathrm{gain\_syn}} I_{\mathrm{w}}}{I_{\mathrm{tau}}} \right) \langle \lambda_{\mathrm{in}} \rangle \, \Delta t \tag{1}$$

where Δt and $\langle \lambda_{\mathrm{in}} \rangle$ are the pulse duration and the average spike frequency respectively.
The synapse block includes a circuit that performs Spike-Timing-Dependent Plasticity (STDP) learning [6]. In these experiments the circuit is biased to achieve a binary weight update. The variation of the weight depends on the time difference between the pre-synaptic (PRE) and post-synaptic (POST) spikes (if PRE before POST, w is driven to the power supply, if POST before PRE, w is driven to ground).

2.1.2 DPI Neuron. The neuron [3] takes advantage of the same principle used for the DPI synapse to obtain a linear behaviour with subthreshold transistors. The circuit is composed of a DPI block and additional circuitry needed to generate the positive and negative feedback necessary for the dynamics that are characteristic of a spiking neuron. To implement the positive feedback, an inverter detects the crossing of the threshold, activating the p-type MOS (pMOS) branch that rapidly charges up the p-type MOSCAP. The positive feedback is then followed by negative feedback, activated by a second inverter that, through an n-type MOS (nMOS) transistor, discharges the V_{mem} node.

The charging of the neuron's membrane depends, similarly to the DPI synapse, on the current reaching the capacitor. As suggested in [3], the speed at which the membrane reaches the threshold impacts the spiking rate (SR) at which it can fire spikes. We can therefore deduce the following:

$$\mathrm{SR} = f \left(\frac{I_{\mathrm{in}} I_{\mathrm{gain\_neu}}}{I_{\mathrm{leak}}} \right) \tag{2}$$

where $I_{\mathrm{in}} = \sum I_{\mathrm{syn}}$.

2.1.3 Synapse Normalisation. The synapse normalisation circuit, interfacing the synapses and neuron, can be seen in Fig. 2 with the label "Normalisation". The part included in the synapse comprises two pMOS transistors. The upper transistor sets the intensity of the normalisation current generated by every synapse, while the second transistor defines digitally whether the normalisation branch of that specific synapse is active or not. Each synapse will activate its normalisation effect only if the weight w is low. The normalisation circuits of every synapse converge into a single node which collects all the currents. Due to this additional circuit, the DPI synapses and DPI neuron are connected through two wires: one transferring the synaptic current and the other one the normalisation current.

Within the neuron, the normalisation circuit is composed of an nMOS branch and a PMOS branch. The nMOS branch collects

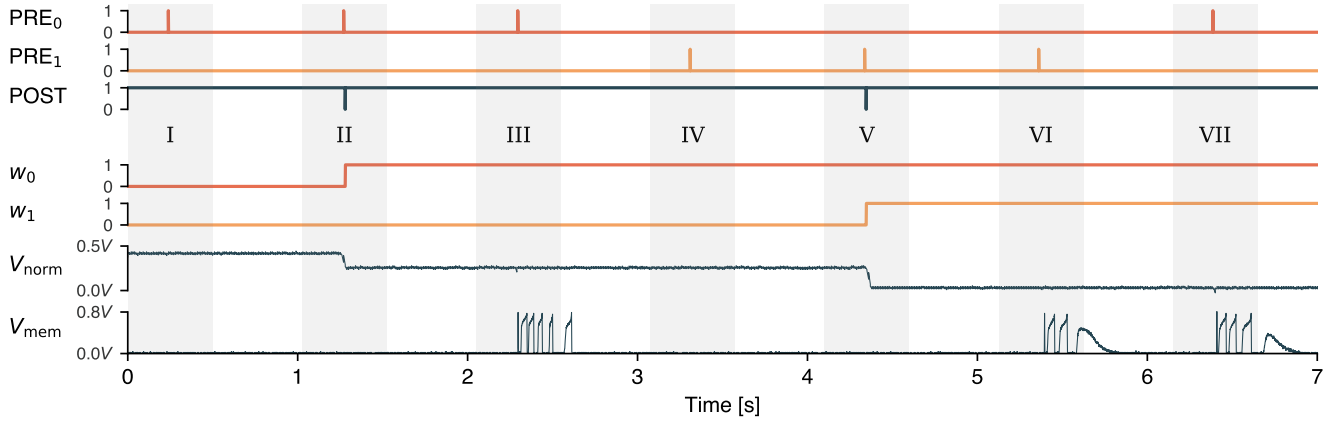

Figure 3: Experimental measurements of on-chip synaptic normalisation. The experiment is divided into 7 phases, described in Section 3.1. PRE_0, PRE_1 and POST are digital inputs from a microcontroller to the circuit. Weights w_0 and w_1 are digital outputs measured using a logic analyser. V_{norm} and V_{mem} are analog voltages measured from the chip using an oscilloscope.

the normalisation current summed from each synapse. Then, due to the diode connected configuration, the neuron gain is set by the received current. The greater the current passing through the normalisation nMOS, the higher the gain of the neuron, according to Equation 2, where this normalisation current would be equivalent to I_{gain_neu}. The pMOS branch is instead an always-on current that biases the gain transistor regardless of the weights. The nMOS collects the contribution of I_{norm_step} and I_{norm_offset}.

Taking into account the circuit operation described above, we can represent the output spike rate (SR) as a function of the I_{gain_neu} current, composed of:

$$I_{gain\_neu} = I_{offset} + \sum_j \overline{w}_j I_{step} \tag{3}$$

Such that:

$$SR_i = f\left(\frac{\left[I_{offset} + \sum_j \overline{w}_j I_{step}\right] \cdot I_w \sum_j w_j}{I_{tau}}\right) \tag{4}$$

In the case where all synapses are in the "off" state, the current flowing in the gain transistor (i.e the transistor connected to the normalisation circuit in the neuron in Fig. 2) is high, so the neuron is easily excited. As synapses become active, their ability to drive the neuron decreases. This produces an effective normalisation of the neuron's activity related to the number of active synapses. Note that the input current increases proportionally with the active weights while the synaptic normalisation decreases. There is also a spurious term where the spiking rate increases proportionally to $I_{offset} \cdot I_w \sum_j w_j$. Note that the input value is binary, where low is 0 and high is I_w.

2.2 Setup for Experimental Measurements

The proposed circuits have been fabricated using the XFAB® 180nm technology. The ASIC comprises a subthreshold Digital to analog Converter (DAC) with 30 channels, and several Operational Transconductance Amplifiers (OTAs) to monitor the analog traces.

We designed a dedicated experimental setup to test the fabricated chip. We used a Cypress FX3® microcontroller programmed with custom firmware. Communication was performed using a custom interface in Python. The program, along with the microcontroller, was used to set the parameters on the chip through an on-chip register chain. Additionally, the program was used to send spikes to the circuits through specific input pads. Oscilloscopes and logic analyzers were used to read out analog and digital signals from the chip.

2.3 Network Simulation

The synaptic normalisation circuits included in the ASIC were designed to test the basic functionality of the blocks. To show the advantage of the synaptic normalization, we designed a software simulation of a network composed of 2 neurons and 18 synapses (9 for each neuron). The software simulation was performed on Cadence Virtuoso®, a state-of-the-art Simulation Program with Integrated Circuit Emphasis (SPICE) simulator with realistic transistor models.

The synapses receive input from a 3×3 matrix of pixels (Fig. 4). Each pixel state is encoded by a single input spike when active and no input spikes otherwise. Two patterns are formed using the matrix of pixels, representing a "O" and a "C"; a large difference in coding level is present between the input stimuli. The task of the network is to learn how to discriminate the two patterns. Specifically, neuron 0 (N_0) should respond when the "O" is presented and neuron 1 (N_1) when the "C" is presented. Each neuron receives the summed synaptic currents that integrate incoming spikes and also the normalisation current that is continuously provided by the synapses to the neuron, changing its gain.

3 RESULTS

3.1 Experimental Measurements

We designed an experimental protocol to demonstrate the ability of the fabricated circuits to properly adapt the neuron's gain in response to changes of synaptic weights. As shown in Fig. 3, the

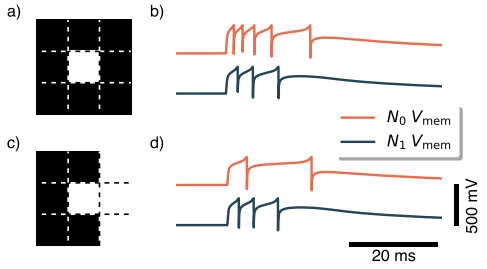

Figure 4: SPICE simulation results of inference for a network of two neurons N_0 and N_1 with 9 normalising synapses each. This can be seen as an example of synaptic normalisation where neurons tuned to inputs with low coding level can be more active than neurons tuned to inputs with high coding level.

experiment starts with both w_{00} and w_{01} set to 0 (phase I). In this configuration, input spikes do not produce synaptic current for the neuron. The normalisation current is at the maximum value $I_{\text{gain\_neu}} = I_{\text{offset}} + 2I_{\text{step}}$ (Eq. 3). The corresponding nMOS gate voltage is labeled in the graph as V_{norm}. The neuron dynamics (trace V_{mem}) do not produce spikes given the lack of synaptic current.

In phase II, the weight of synapse 0 is changed as a result of pre-post spike pair stimulation with a delay of 10 ms (using the STDP circuit [6]). Also in this case the neuron does not produce spikes because the synaptic weight was still zero at the arrival of the pre-synapstic spike. Nevertheless, the input stimulation makes the synaptic weight potentiate (change from 0 to 1). Therefore, $I_{\text{gain\_neu}}$ becomes $I_{\text{offset}} + I_{\text{step}}$, effectively reducing the excitability of the neuron. A PRE_0 spike is then sent again, causing the neuron to spike with a given spike count (phase III).

The same protocol is then repeated for synapse 1 (phase IV, V and VI). The final outcome is different in this case (compare phase III with phase VI/VII), given that both w_{00} and w_{01} are potentiated, $I_{\text{gain\_neu}}$ is only as large as I_{offset}, the lowest possible value. For this reason, the neuron responds with fewer spikes to an input pulse applied to synapse 1 (phase VI) and synapse 0 (phase VII). In particular, the response to a pulse on PRE_0 shows that the response of synapse 0 is altered when w_{01} is potentiated (compare phase III and VII).

3.2 Network Simulation Result

In order to show the potential application of the synaptic normalisation circuit, a bigger network composed by 18 synapses and 2 neurons has been simulated using Cadence Spectre®. To emulate the pixel representation, the simulation employs 9 different voltage generators that create a single spike at the correct positions. For pattern "O" and "C" out of 9 generators, 8 and 5 were active, respectively (Fig. 4).

To test inference capabilities of the network, synaptic weights are pre-programmed so that neuron N_0 is stimulated by all active pixels of pattern "O" and neuron N_1 by those of pattern "C". In the experiment (Fig. 4), the synapses are activated by input spikes representing the pattern "O" and then the pattern "C". Given that pattern "C" fully overlaps with pattern "O", the network would fail to distinguish them without synaptic normalization because the two output neurons would be equally active upon presentation of pattern "C". Synaptic normalisation reduces the spiking activity of neuron N_0 in response to pattern "C", because only a subset of its active synapses are stimulated. Instead, neuron N_1 can strongly respond to pattern "C" thanks to the fact that all its active synapses are stimulated. Therefore an online comparison of the output spike count or instantaneous firing rate of the two neurons leads to a correct classification of the input pattern. This can be explained considering Equation 4. From that we can calculate that N_0 and N_1 have a normalisation current respectively of $I_{\text{offset}} + I_{\text{step}}$ and $I_{\text{offset}} + 4I_{\text{step}}$, while they have an input current of $8I_w$ and $5I_w$.

4 CONCLUSIONS

In this work, we presented a novel CMOS circuit that implements synaptic normalisation in an analog hardware SNN in combination with STDP. The effective functionality of such a technique was validated through experimental measurements performed on ASIC and assessed through circuit simulations. The circuit, with its minimum footprint (3060 μm^2, without DAC and OTA periphery) can be easily scaled up to hundreds of synapses per neuron, varying the step and offset current given by every weight variation.

For future work we plan to extend the circuit to facilitate analog weights. The circuit is agnostic to the learning rule implemented and can be used in conjunction with learning rules for which the neuron requires knowledge of the weights of input synapses [4].

ACKNOWLEDGMENTS

The authors would like to acknowledge Philipp Klein, Nicoletta Risi, Ton Juny Pina and Maxime Fabre for assisting with the research. Picture Credit in Fig. 1 to Thorben Schoepe and O.R. This work has been supported by EU H2020 projects NeuTouch (813713), BeFerroSynaptic (871737) and MANIC (861153). Additional funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): Project MemTDE Project number 441959088 as part of the DFG priority program SPP 2262 MemrisTec Project number 422738993; Project NMVAC Project number 432009531. The authors would like to acknowledge the financial support of the CogniGron research center and the Ubbo Emmius Funds (Univ. of Groningen).

REFERENCES

[1] Chiara Bartolozzi and Giacomo Indiveri. 2007. Synaptic dynamics in analog vlsi. *Neural Computation*, 19, 10, (Oct. 2007), 2581–2603.

[2] Joseph M. Brader, Walter Senn, and Stefano Fusi. 2007. Learning real-world stimuli in a neural network with spike-driven synaptic dynamics. eng. *Neural Computation*, 19, 11, (Nov. 2007), 2881–2912.

[3] Elisabetta Chicca, Fabio Stefanini, Chiara Bartolozzi, and Giacomo Indiveri. 2014. Neuromorphic electronic circuits for building autonomous cognitive systems. *Proceedings of the IEEE*, 102, 9, 1367–1388.

[4] Giorgia Dellaferrera, Stanisław Woźniak, Giacomo Indiveri, Angeliki Pantazi, and Evangelos Eleftheriou. 2022. Introducing principles of synaptic integration in the optimization of deep neural networks. en. *Nature Communications*, 13, 11, (Apr. 2022), 1885.

[5] Charlotte Frenkel, David Bol, and Giacomo Indiveri. 2021. Bottom-up and top-down neural processing systems design: neuromorphic intelligence as the convergence of natural and artificial intelligence. (2021).

[6] Giacomo Indiveri, Elisabetta Chicca, and Rodney Douglas. 2006. A vlsi array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity. *IEEE Transactions on Neural Networks*, 17, 1, 211–221.

[7] Walter Senn and Stefano Fusi. 2005. Convergence of stochastic learning in perceptrons with binary synapses. *Phys. Rev. E*, 71, (June 2005), 061907, 6, (June 2005).