Neuromorphic Computing

SEMINAR REPORT

Submitted to:

Department of Electronics and Communication Engineering

Submitted by:

Name — Vedika Jain
Roll Number — 21BEC018

Under the Supervision of:
Prof. Sajad Ahmad Loan

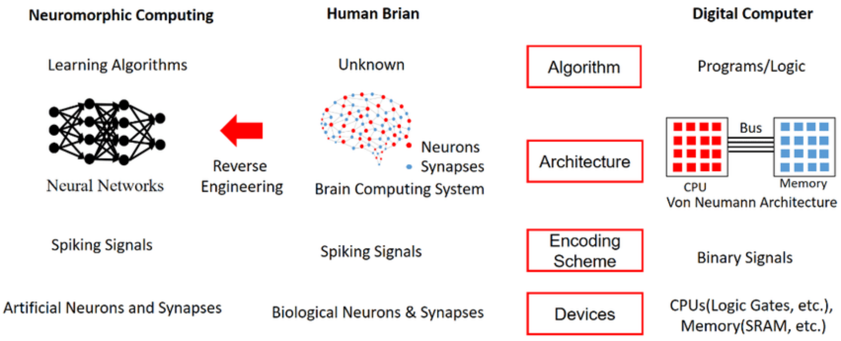
Faculty of Engineering and Technology

Jamia Millia Islamia, New Delhi

List of Contents

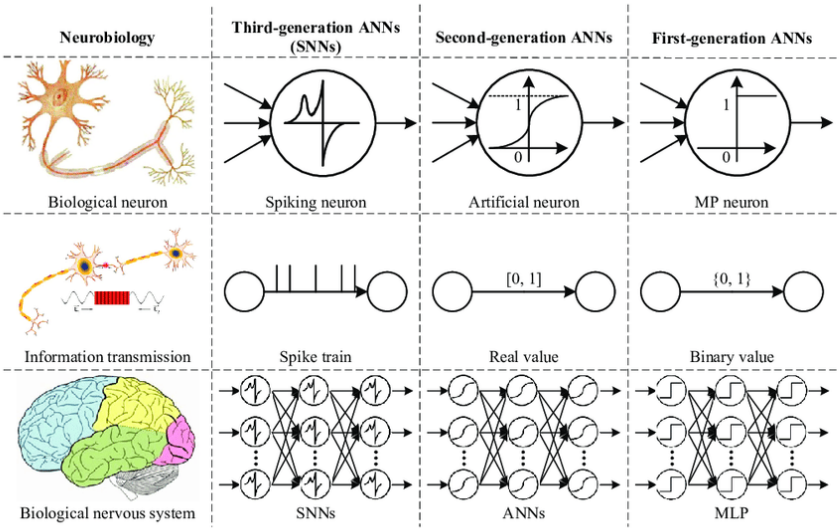
Introduction

Traditional computing faces challenges in tasks such as pattern recognition, sensory processing, and real-time interaction, which are effortlessly performed by the human brain. AI is best run on hardware which is similar to the abstraction. The neurons in the brain do not discriminate between computation and memory. The connections between neurons keep evolving, lending us the ability to learn and unlearn. We don't think in binary, nature is analog. Only the active parts of the brain consume energy. The present models are very power intensive and inefficient. If we are to harness the power of AI, we need hardware which is high on speed, low on power consumption, and nimble with computations. The present von Neumann architecture has hit these roadblocks and can’t support AI models for long.



A comparison of von Neumann architecture based systems and the brain. Neuromorphic systems can give us the best of both worlds.

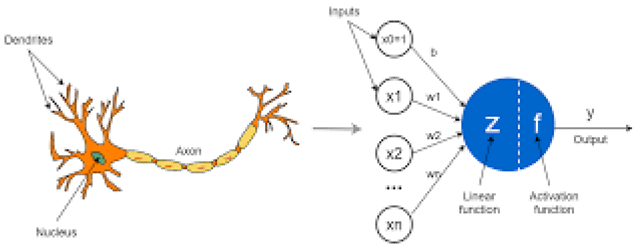
Neuromorphic computing, inspired by the brain's efficiency and parallelism, aims to overcome these limitations. Neuromorphic computing seeks to mimic the structure and function of the human brain using electronic circuits. A key idea in neuromorphic computing is that of Artificial Neural Networks (ANNs). ANNs are computational models inspired by the structure and function of biological neural networks.



According to their computational units, ANN models can be divided into three different generations.

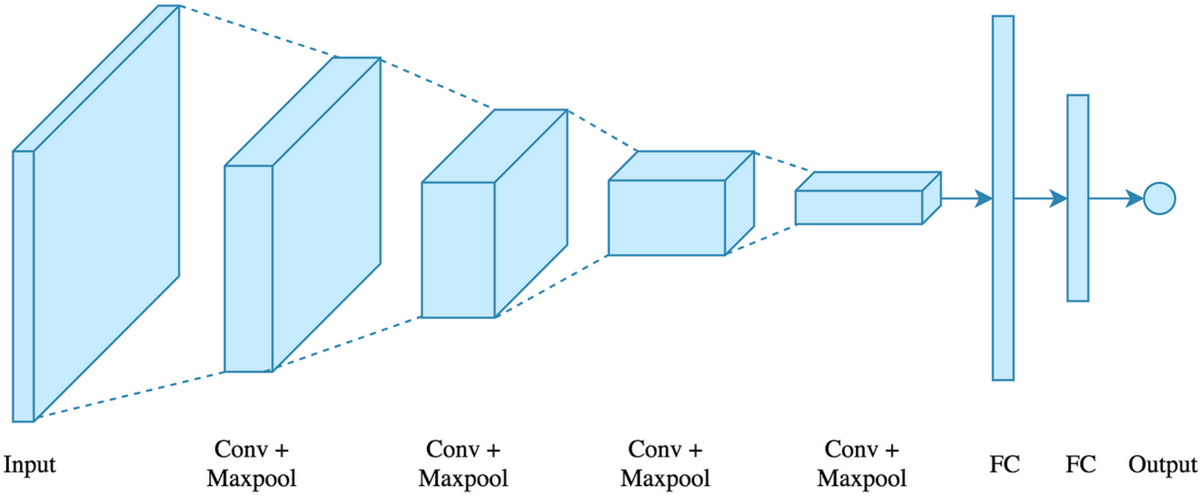
The first generation ANNs used McCulloch Pitts neuron, which simply checked if the input is higher than a certain threshold. They were not powerful and were based on old technology, and culminated into the AI winter.

Present systems make use of the second generation ANNs. They are based on what is called the ‘artificial neuron’, which is nothing like its biological counterpart. It scales and biases the input, and applies a mathematical function to it. The rise of this generation was facilitated by GPUs, more data, and new algorithm optimisation techniques.



Biological Neuron vis-à-vis Artificial Neuron

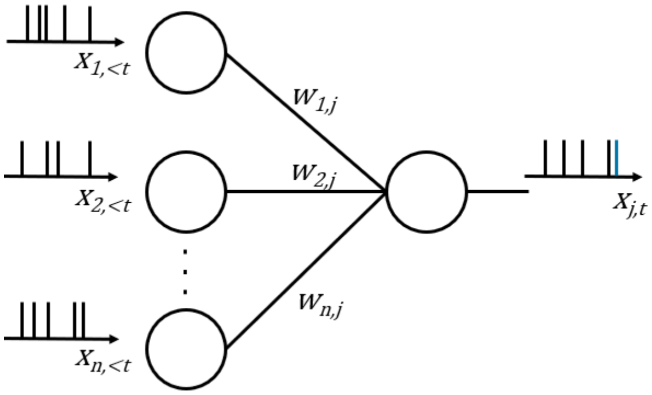
The most popular are the Convolutional Neural Networks (CNNs). CNNs allow us to work with big data while keeping the number of trainable parameters reasonable. This is achieved by using convolutional filters to operate on multiple inputs at once, and pooling layers to decrease the size of activations. It is because of CNNs that computer vision and large language models (LLMs) became a reality. However, CNNs are computationally expensive and hence are not scalable.



General blueprint of a CNN

Technical Background

Third-generation ANNs are Spiking Neural Networks (SNNs) that use biologically plausible spiking neurons as the basic computational units. The neuron model used depends on the application. Neural information in the spiking neuron is transmitted and processed by precisely timed spike trains. Efficacy of SNNs depends on the spike train encoding used. A spike can be represented as a “glitch”, which can be obtained when different bit streams travelling along paths with different propagation delays meet.

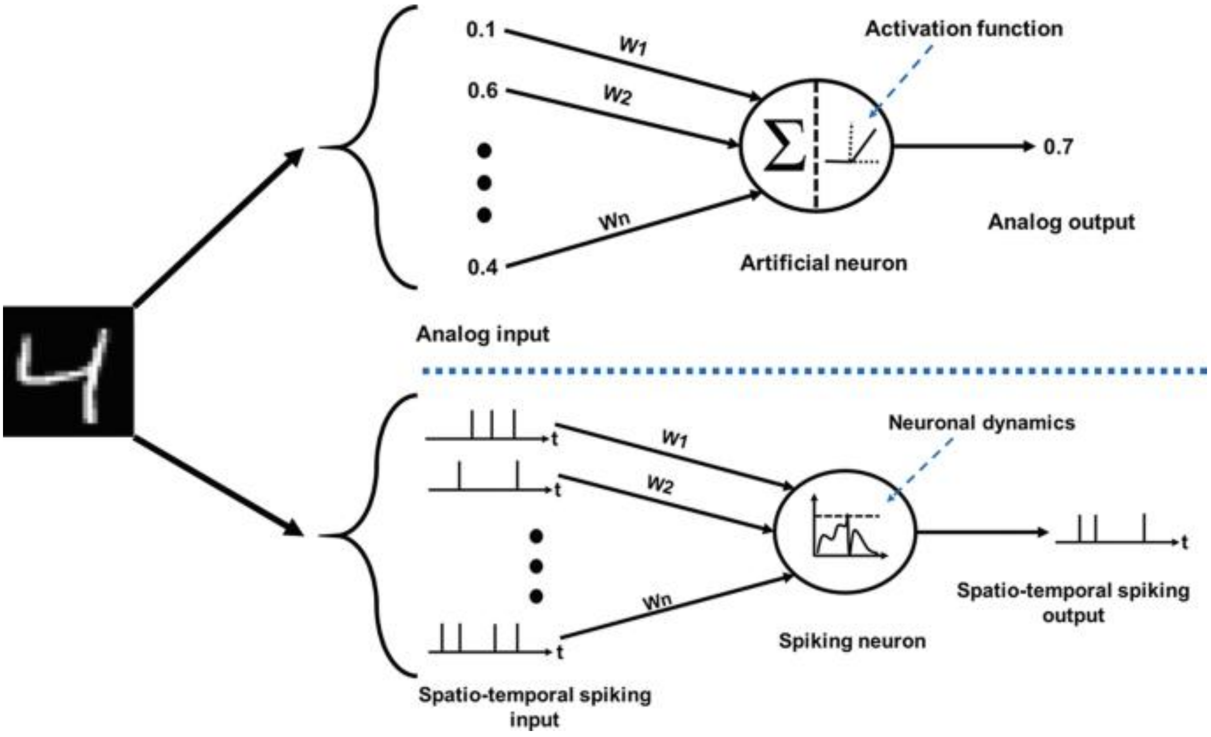


A sample SNN

Updating of synaptic weights in SNNs occurs on the basis of spike-timing-dependent-plasticity (STDP). It is based on Hebb’s rule, which is often colloquially quoted as, “Neurons that fire together, wire together.” Hebb’s rule overlooks the timing of the pre-synaptic and post-synaptic fires, which is remedied by STDP. STDP adds another condition, “Those who fire out of sync, lose their link.”

Compared with the first- and second-generation ANN models, SNNs can describe the real biological nervous system more accurately, so as to achieve efficient information processing. Spike trains make them more power efficient compared to ANNs, which work on analog data. ANNs can’t capture time and space dependencies present in the data effectively. Therefore, SNNs have more powerful information representation ability than ANNs.

The state of a spiking neuron is represented by differential equations, whereas ANNs use activation functions for this purpose. SNNs reinforce causality using STDP; Hebb’s rule used by ANNs ignores this. SNNs are fine-tuned using mentalities; ANNs are optimised using objective functions, also called loss functions. SNN execution is clock or event driven; ANN execution flows sequentially.



ANN and SNN processing an MNIST sample

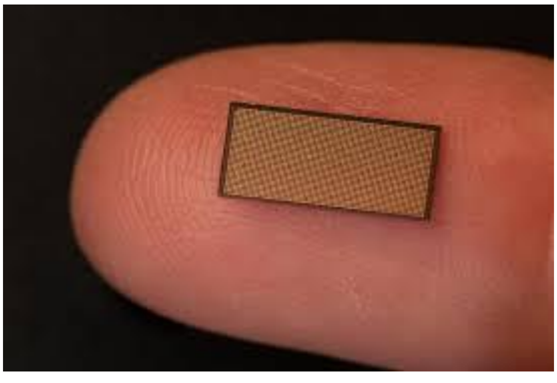
With SNNs, there exists a trade off between accuracy and simplicity. Biologically accurate neuron models like the Hodgkin-Huxley model are complex to implement and are computationally expensive, whereas simpler models like the integrate-and-fire model aren’t accurate. A balance can be achieved by using approximated models like the leaky integrate-and-fire (LIF) neuron or the Izhikevich neuron.

Existing Work

Neuromorphic computing is still a novel field, with constant research underway and innovations being broken every day. A lot of our research is focussed on understanding the biological brain, which influences our system design too. As our understanding of the brain widens, the architecture is also modified accordingly. Academia mainly looks for new opportunities in this domain, whereas industry giants are working on powerful neuromorphic chips. Two forerunners are IBM and Intel.

IBM, credited to many firsts of technology, built the TrueNorth chip in 2014. Its successor, NorthPole, was released last year. NorthPole is 4000 times faster than its predecessor, and can run tasks like ResNet-50 or Yolo-v4 image recognition about 22 times faster, with 25 times less energy and 5 times less space, when compared to GPUs (like Nvidia’s V100) which use the same 12-nm node process that it was fabricated with. It even outperformed chips made with more advanced nodes. For example, compared with Nvidia’s H100 GPU, implemented using a 4-nm node, NorthPole was five times more energy efficient. It includes 224 MB of RAM and 256 processor cores and can perform 2,048 operations per core per cycle at 8-bit precision, and 8,192 operations at 2-bit precision. It runs at between 25 and 425 MHz. It can be integrated into systems too as it appears as an active memory chip from outside the device.

Intel made the Loihi chip in 2017, followed by Loihi 2 in 2021. Advances in Loihi 2 provide up to 10 times faster processing, up to 15 times greater resource density with up to 1 million neurons per chip, and improved energy efficiency. Early evaluations suggest reductions of over 60 times fewer ops per inference compared to standard deep networks running on the original Loihi without loss in accuracy. Fully programmable neuron models and generalized spike messaging open the door to a wide range of new neural network models that can be trained in deep learning. It improves support for advanced learning methods, including variations of backpropagation. Greater programmability also allows a wider class of difficult optimization problems to be supported. This expands the scope of adaptation and data efficient learning algorithms that can be supported by low-power form factors operating in online settings. Loihi 2 addresses a practical limitation of its predecessor by incorporating faster, more flexible, and more standard input/output interfaces.



Intel’s Loihi 2 placed on a fingertip for scale demonstration

The Human Brain Project (HBP), funded by the European Union, boasts of two flagship supercomputers: SpiNNaker and BrainScaleS. SpiNNaker (Spiking Neural Network Architecture) is a massively parallel, many-core supercomputer architecture designed by the Advanced Processor Technologies Research Group at the Department of Computer Science, University of Manchester.

The Tianjic chip seeks to unify the divergent neuroscience and computational approaches towards the development of artificial general intelligence. It adopts a many-core architecture, reconfigurable building blocks and a streamlined dataflow with hybrid coding schemes, and can not only accommodate computer-science-based machine-learning algorithms, but also easily implement brain-inspired circuits and several coding schemes.

Conclusion

With Moore’s law being transformed into more than Moore, the scientific community is exploring alternate avenues for faster, cheaper, and more efficient computing, and Neuromorphic Computing is found to be one of the viable replacements of the present computing paradigm. Neuromorphic computing can be realized efficiently by utilizing Spiking Neural Networks (SNN). It aims at realizing the architecture and performance of a brain in silicon. The brain, being the most efficient system present, processes complex information much faster than any existing computer. In recent years, the popularity and application of spiking neural networks have increased considerably.

SNN is prominently used in many applications such as event detection, classification, speech recognition, spatial navigation, and autonomous motor control. It has demonstrated its effectiveness in detecting analog signals from sensors; designing controllers for autonomous robots; performing detection and recognition tasks; processing cortical data, and tactile form-based recognition.

With the advent of autonomous robots and self-driving vehicles and due to the rise in the real time applications of embedded systems, it has become imperative to realize machine learning models on compact and energy-efficient platforms.

The existing neural network models are computationally intensive and require huge memory for their realization, therefore making them unsuitable for real time and energy-efficient applications. Although SNN has been realized in Applications Specific Integrated Circuits (ASICs) such as SpiNNaker, BrainScaleS, SyNAPSE, Neuropipe-chip, etc., their objective is mainly to provide a solution for large scale simulations rather than for low power embedded applications.

Conventionally, Field Programmable Gate Arrays (FPGAs) are utilized for the validation of electronic systems as well as in the implementation of time-critical systems. FPGAs are also well-suited for providing a low power solution for massively parallel and computationally less complex models. Although parallelism can be achieved using Graphics Processing Units (GPUs) as well, FPGA implementations are advantageous where power consumption is an issue.

Considering the arguments presented here, therefore, the Minor and Major Projects would work towards implementing an SNN on an FPGA.

References

[1] Wang, Xiangwen & Lin, Xianghong & Dang, Xiaochao. (2020). Supervised learning in spiking neural networks: A review of algorithms and evaluations. Neural Networks. 125. 258-280. 10.1016/j.neunet.2020.02.011.

[2] Fang, Haowen, et al. "Encoding, model, and architecture: Systematic optimization for spiking neural network in FPGAs." Proceedings of the 39th International Conference on Computer-Aided Design. 2020.

[3] Gupta, Shikhar, Arpan Vyas, and Gaurav Trivedi. "FPGA implementation of simplified spiking neural network." 2020 27th IEEE International Conference on Electronics, Circuits and Systems (ICECS). IEEE, 2020.

[4] Ju, Xiping, et al. "An FPGA implementation of deep spiking neural networks for low-power and fast classification." Neural computation 32.1 (2020): 182-204.

[5] Lent, Ricardo. "Evaluating the cognitive network controller with an SNN on FPGA." 2020 IEEE International Conference on Wireless for Space and Extreme Environments (WiSEE). IEEE, 2020.

[6] Izhikevich, Eugene M. "Simple model of spiking neurons." IEEE Transactions on neural networks 14.6 (2003): 1569-1572.

[7] Cassidy, Andrew S., et al. "11.4 IBM NorthPole: An Architecture for Neural Network Inference with a 12nm Chip." 2024 IEEE International Solid-State Circuits Conference (ISSCC). Vol. 67. IEEE, 2024.

[8] Orchard, Garrick, et al. "Efficient neuromorphic signal processing with loihi 2." 2021 IEEE Workshop on Signal Processing Systems (SiPS). IEEE, 2021.

[9] Khan, Muhammad Mukaram, et al. "SpiNNaker: mapping neural networks onto a massively-parallel chip multiprocessor." 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE, 2008.

[10] Pei, Jing, et al. "Towards artificial general intelligence with hybrid Tianjic chip architecture." Nature 572.7767 (2019): 106-111.

[11] An, Hongyu & Bai, Kangjun & Yi, Yang. (2018). The Roadmap to Realizing Memristive Three-dimensional Neuromorphic Computing System.

[12] Katamreddy, Sukumar & Doody, Pat & Walsh, Joseph & Riordan, Daniel. (2018). Visual Udder Detection with Deep Neural Networks. 166-171. 10.1109/ICSensT.2018.8603625.

[13] Acharya, J., Basu, A. (2023). Neuromorphic Spiking Neural Network Algorithms. In: Thakor, N.V. (eds) Handbook of Neuroengineering. Springer, Singapore. https://doi.org/10.1007/978-981-16-5540-1_44