Name: URVI PATEL UID:22BIT034
Subject: IOT Date: 26/1/2024

Neuromorphic Computing

Introduction

The von Neumann architecture has long been the computational paradigm, followed even by neural networks claiming to mimic the brain. However, in reality, our brain doesn't have separate sections to store memories and perform processing, which is why it takes much less space, energy and time to process information than a computer, or even an ANN. Enter Neuromorphic Computing. This takes inspiration from the voltages ruling our brain, namely membrane potential and action potential, to create Spiking Neurons. Basically, when input voltage to a neuron exceeds its membrane voltage/potential, it gets triggered and fires off action potential to the next neuron through the synapse. This leads to a transmission of signals through a chain of neurons. These brief action potentials are called "spikes" and closely resemble our natural neurons' behaviour of inactivity or activity based on potential. Such a practise is known as Event-driven processing. Spiking neural networks (SNN) also borrow from brain plasticity in the concept of Spiking Time-Dependant Plasticity (STDP) which means they adapt and learn flexibly, and hence faster than today's ANN models. Neuromorphic computing makes an extensive use of memristors, that stand in for synapses in carrying signals between neurons. These have high and low resistance states and switch between the two on the application of a certain voltage for an apt amount of time. Switching from high resistance to low resistance allows the current to pass through, thus mimicking the spiking of information signals from one neuron to another on application of action potential. This leads to greater learning rates in ML too. IOT is an immediate beneficiary of Neuromorphic computing, with its constant need of being energy-efficient. Moreover, with the constantly increasing data, there is a rampant need for real-time processing. Neuromorphic computing, with its capabilities of speed and abysmal energy requirements, rises to the expectation of adaptive learning on the edge. This is much like our brain constantly re-adjusts to changing environments and doesn't necessarily need to be trained as in conventional neural network models. The developments being made in vision, auditory, olfactory sensors prove that neuromorphic computing is in need of greater research for its progress in IOT, as it rejects the conventional frame-based processing. Instead, it works to take the image in its spatio-temporal context as a whole, leading to event-based visual processing. Further uses of SNNs and memristors can be found in smart manufacturing, smart engines that help in controlling carbon output for the environment's health and navigation systems. Commercially, Dynamic Vision Sensors are seeing a fair amount of success and so are neuro-inspired olfaction units like Cyranose which help detect target chemicals among a slew of background odours. Initiatives like the Human Brain Project are aiming to provide bigger platforms to work on Neuromorphic models, which have been extensively described in the following summary report. Further, companies like Intel's Loihi, IBM's TrueNorth and Neuralink are scrupulous in studying the "potential" of integrating neurobiology in computing. Neurosonics, a company specialising in medical devices, has developed a proof-of-concept with one of the Neuromorphic architectures called FPGA (Field-Programmable Gate Array) to help treat neurological diseases through the brain or spinal cord. Through this summary report, I have attempted to understand Neuromorphic Computing and its applications in IOT. For accomplishing this, I have studied 6 papers and written their literature reviews.

Energy and Area Efficiency in Neuromorphic Computing for Resource Constrained Devices

Gangotree Chakma, Nicholas D. Skuda, Catherine D. Schuman, James S. Plank, Mark E. Dean, Garrett S. Rose

23rd May, 2018

{memristor, sparse RNN, NIDA, switching time, STDP, IAF neuron, CMOS}

IOT as an emerging field uses resource-constrained devices with ML capabilities, like autonomous drones with simple neural nets that learn and adapt to small changes in environment. Such simple drones are effective in searching disaster victims. But, such low power, real time processing systems are challenging to design. Cascading neural nets have been shown to distinguish MNIST characters using IOT devices, with small error rate. A System-on-Chip (SoC) use of CNNs reduces the number of threads, and thus, power to operate DL apps. Neuromorphic computing thus has much scope in working with constraints to build efficient DL systems for IOT. Here, sparse RNNs are studied as they save much energy compared to CNNs. An evolutionary optimisation technique is introduced to make compact SNNs. Sparse RNNs are used since they have less connections to make a neuro-inspired dynamic architecture (NIDA), oft simulated as a HL brain-like structure. This means less connections, less space and thus, less power is used to achieve brain-like processing of information on Iot devices. In addition, NIDA-based memristive dynamic adaptive neural network architecture (mrDANNA) is also being developed. Data analysis of large volumes of sensor data is done by DNNs. To avoid a computational load on edge computing, the data is encrypted and compressed to reach data centre before processing and responding to device aptly. This leads to a trade-off between low power and reliability. Adding computational power to IOT devices may not be feasible, hence the call for small, low-power devices.

Leon O. Chua introduced memristor, a 2-terminal nano-scale non-volatile device, whose resistance is modulated by the magnitude of voltage and time for which voltage is applied. The resistance achieved is bounded by low resistance state (LRS) and the high resistance state (HRS), depending upon switching material, process condition, noise and environment factors. Being small, they are used in area-efficient designs and large LRS devices have lower power consumption, making them ideal to design energy efficient neuromorphic chips. Use of switching materials is beneficial to implement synaptic circuits. Nanoscale hafnium-oxide (H f Ox) memristors fabricated within a 65 nm CMOS process are used to develop the model. When voltage> 1V (threshold) is applied, device switches from HRS to LRS. The switch from LRS to HRS occurs when negative voltage is applied <-0.7V. both these voltages must be applied minimum “switching time” (typically 10 − 100ns) in order to switch fully between the two extreme resistance states, HRS and LRS. These switches can also be achieved by nudging the resistance by short, ns pulses. The variety and range in resistance helps denote synaptic weights in the SNN model considered. The changing of resistive values by controlled pulses helps encourage on-line learning mechanisms like Spike Time-Dependant Plasticity (STDP). To make the synapses, non-volatile memristors are used to consume low energy and store synaptic weights. The energy per spike consumed by Hf Ox is given in idle state, active phase and learning phase (potentiation and depression on firing neuron), and the

values are abysmally low. Each of these synapse states consumes energy depending on the type of memristive device used and the peripheral control circuitry. (the circuitry involved to handle neurons in larger system for communication, interrupt handling etc). mixed signal neurons are used with very few transistors, considering area- and energy-efficiency. The capacitor size though a concern can be solved by using memristors that control incoming current to neuron. The neurons operate in 3 phases: accumulation, idle and firing phases. Here the accumulation energy refers to the energy consumed while accumulating incoming charge/spikes from the synaptic weights. The idle energy is low since the neuron's functionality is inactive except peripheral circuitry. Energy is consumed while firing spikes, not while learning. This energy involves post-neuron fire when accumulated charge>threshold of neuron. The shape of firing spike is vital for energy consumption as generated firing pulses are fed into the neuromorphic core (a brain-like processing unit). mrDANNA is the mixed signal neuromorphic architecture, that is area- and energy-efficient while including synapse and neuromorphic model. Several synapses and neurons make up a mrDANNA core, consisting of energy-efficient memristive synapses and analog IAF (integrate and fire) neuron. 36 such cores are placed at right side of architecture, digitally implemented on left side. Connections with outside signal are fully digital and integration with core is analog, hence inter-core connections are digital and intra-core is analog. Apart from being efficient in area and power, it is easier to implement synapse and neuron model in neuromorphic cores for applications like anomaly detection, classification and control.

An application is an autonomous robot's navigation system, with the robot having fixed size, energy and storage resources. NeoN robot with mrDANNA chip is considered. Output spikes activate motor controller and input spikes are produced by sensors (here LIDAR sensors are used). The robot is supposed to avoid obstacles while exploring largest area possible in an unknown environment. Networks with EO are used to simulate this task. The primitive simulation does well in training the network by EO in another neuromorphic architecture called DANNA, then used to control robots in simulated environments. The example network generated using the EO to operate the robot uses 31 neurons (9 input neurons, 4 output neurons, and 18 hidden neurons) and utilizes 119 of the synapses for communication, twice as many as the DANNA implementation. The LIDAR information is fed in 5 inputs while 2 inputs are used for robot's limit switch information and a bias and random value for activity. A single sparse neuron layer handles input and decision making, with many recurrent connections. This approach, instead of traditional feed-forward neural nets is what makes networks smaller. The performance of network on physical hardware is shown by measuring various types of activities on network. In a system with 31 neurons and 119 synapses, the neurons collectively generate an average of 4425 spikes per second. The core logic uses power of 142.7 micro-watts. The non-trivial costs of producing spikes from input sensors or spikes to output signals to motor controller are not considered. The generated network is sparse, with some heavily connected nodes, with outdegrees up to 13 and indegrees up to 11. The connectivity is primarily directly between input and outputs neurons (including many input-to-input connections), with the hidden neurons usually having relatively few synapses.

The authors emphasise use of neuromorphic computing in area- and power-efficient IOT. The result of the study using mrDANNA for making an autonomous robot navigation system with evolutionary optimisation, led to a comparatively sparse and small DL network which takes up 142.7 micro-watts, thus fulfilling both area- and power-efficiency. The authors feel that

there is a lot of research to use neuromorphic computing in data communications and data analysis. The use of on-chip plasticity mechanisms leads to self-healing devices for inaccessible IOT devices. The research in Big Data continues while ensuring energy and area efficiency in largely connected IOT devices.

The Human Brain Project and Neuromorphic Computing

Andrea Calimera, Enrico Macii, Massimo Poncino

17th Oct, 2013

{perceptron, accelerators, FPGA, SpiNNaker, Network-on-Chip, Leaky IAF}

Carver Mead coined “neuromorphic computing” in 1980s. it stood for electronic systems operating on similar computation as our nervous system, but the concept has widened significantly for including computation and neural systems. Originally used to simulate the working of the brain, it is now used to overcome current technological limitations of area and energy-efficiency. This leads to the study of architectures that can mimic brain like operations while reducing volume and energy per operation. The old and new interpretations are similar in that they both exploit a particular chip design feature like computational speed and automated/ consolidated design flows. Both deal with energy consumption as a problem. In contrast to programmable von Neumann architectures, neuromorphic architectures require highly parallelised, learning-cantered approach. Such circuits are designed on CMOS devices and use highly standardised, automated tools and flows for design.

Earlier attempts to design electronic models of neural circuits consisted of “brain-centred” areas like perceptron and retinas. The properties of electronic devices, circuits and systems started to be exploited in the 80s for simulating brain models. The development of designs, implementations and prototype chips in the last 20 years have impacted NC’s progress in the common goal of mimicking the brain within energy constraints. The various solutions, though difficult to fully cover, can be grouped in 2 main classes: emulative and simulative.

Emulative approach uses noisy electronic components, like atom-scale sized transistors to physically emulate neural models. The resulting circuits are called neurochips and they exploit the non-linear current characteristics of silicon transistors to copy the electrochemical functions of NNs. They can be classified as analog or digital. Analog circuits were used by Mead and are still used by ongoing research as they integrate bio-inspired electronic sensors with analog circuits. Being compact and low-energy consuming, they can perform neural functions like integration and summation of currents and charges. But they are sensitive to noise, variations in physical parameters like doping, oxide thickness, and are difficult to design as technology nodes shrink, leading to a greater implementation cycle (more time to implement). Digital simulative solutions are more reliable, computationally powerful and faster to design due to CAD tools pf large VLSIs. The circuit size, due to lack of elementary functions (integration) as were in analog, becomes huge.

Neuro-computers, as in the simulative solutions focus on imitating the neural models rather than accurately copying neural signals. The cheap, reliable, highly available ICs are used to abstract the innards of the brain’s working and produce large-scale structure. This is used to focus on how sensors take input, make connections and produce a motor output. “Accelerator boards”, “programmable arrays” and “general-purpose” are the 3 types of neurocomputers, that are implemented to fulfil the evolving needs of neurocomputers. Accelerators are used to speed up the processing in an existent neural model. Being the 1st kind of NC hardware, these

were cheap, user-friendly and could be connected to a PC by slots. Usually based on ANNs, DSPs are also used for faster signal processing, up to 10x the normal simulation software. Being specialised, they lack flexibility and are harder to design for adaptations. To counter this, field programmable gate arrays – FPGAs are a better solution as they encourage real time reprogrammability. But these face power and routability issues, stopping them from being used to scale a neural model to bigger sizes. The 3rd approach relies on general purpose processes and software, offering enough programmability to study different neural models. They are the least accurate, but they provide a wide array of NC models to scale, which helps refine existing models and optimises other NC hardware like accelerators and neurochips. The software overhead and the "flight time" to exchange data between CPU and memory is heavy on the performance. Implementations range from low-cost to complex architectures with parallel i/o lines for performing discrete Fourier transforms. Recently, multicore architectures are leveraged, as in the SpiNNaker platform. There is no best out of the 3 approaches as all work to understand the human brain.

The Human Brain Project (HBP) provides certain services to use hardware and software for operations, configurations and data analysis in context of NC research. This is done through Neuromorphic Computing Platform (NCP) and the services correspond to two neuromorphic computing systems (NCSs), i.e., specific and custom-designed neuromorphic "hardware" systems which are the project's outcome.

The 1st NCS is based on the European FACETS project, called NM-PM (Physical Model). It combines local, analog synaptic communications with binary, async, continuous-time spike communication. These use 20000000 plastic synapses and 200000 biologically realistic neuron models on a single 8inch silicon wafer. This is a mixed VLSI implementation on a standard CMOS 180nm processor. Custom designed analog circuits using async spike exchanges to communicate are used for local synaptic computations. Plasticity mechanisms, complex neuron models, spike frequency adaptation and bio inspired firing models are used. SRAMs and analog circuits are used in the non-von Neumann architecture with decentralised memory. The silicon wafers are stacked to convey computational efficiency. FPGAs are interfacing the digital portion and the analog circuits. This digital portion is contained in motherboard where the silicon stacks are connected. This system has a key feature of not executing programmed code, but evolving by the physical properties of devices, thus following the principles of neuromorphic hardware. The network description language, PyNN, provides portable access to software simulators and neuromorphic systems. The evolution of this NM-PM paradigm is described following 3 models, over the course of 3 years and aiming to emulate a significant fraction of the human brain.

The 2nd NCS is based on heavily parallel multicore ARM processors, hence named NM-MC (Multi Core) or SpiNNaker. 18 ARM cores share 128MB local memory. A single chip simulates 16,000 neurons with 8 million plastic synapses running in real time within an energy budget of 1W. One of the cores is a monitor processor and the others are fascicle cores that models a group of up to 1000 neurons. Spike events communicate through internal chip routing table, called Network-on-Chip. Multiple interconnected chips use 6 bidirectional transmit and receive interfaces to communicate with neighbouring chips. This creates an effective brain-like structure which can be used for all computation types. Software running on it is thoroughly developed for OS compatibility and internode communications. There is no single algorithm but many such algos, neural patterns and connectivity that NM-MC can

be used to evaluate. 2 versions of NM-MC are expected to be developed over 3 years to increase its computational complexity. The NM-PM uses exponential integrate-and-fire (AdExp) neuron model and SpiNNaker, tailored to study the SNN in context of engineering applications, is optimised for Leaky IAF models. The NM-MC is not aiming for biological accuracy, but for increasing a computational perspective.

This paper emphasises on the importance of neuromorphic computing in increasing the feedback that improves neural models developed by neuroscientists. This is done through 2 paradigms proposed by The Neuromorphic Computing Platform in the Human Brain Project.

Name: URVI PATEL UID:22BIT034
Subject: IOT Date: 26/1/2024

Challenges and Perspectives in Neuromorphic-based Visual IoT Systems and Networks

Maria Martini and Nabeel Khan, Yin Bi and Yiannis Andreopoulos, Hadi Saki and M. Shikh-Bahaei

09 April 2020

{DVS, spatio-temporal resolution, up linking, event rate, Active Pixel Sensing, event asynchronicity, event sparsity}

IOT is seen as a pioneering technology in that it provides protocol stacks like MQTT or 6LoWPAN for the transfer of sensor generated data to cloud services for analysis and classification/ recognition. A factor contributing to rise of IOT deployments is the high frame rate achieved by surveillance drones for visual sensing at a low power. High bandwidth requirements turn attention to bio-inspired visual systems, which are highly async and highly speedy. Human vision system detects motion and reflectance via photoreceptor cells (cones & rods) and the visual cortex adds information. Neuromorphic/ Dynamic Vision Sensors (DVS) devices output data as coordinates and timestamps of reflectance events and are triggered when the log of the pixel's intensity value of a planar CMOS sensor goes beyond a threshold. Compared to a conventional frame sensor, DVS captures the binary triggering of fast motion events with less power (only 10-20 mW, rather than 100s of mW) and an increase in speed up to 700-2000 fps. They are commercially available too, unlike compressive sensors, and therefore used in many apps like surveillance of high-speed events or visual stimulation implants for the impaired. A framework called Internet of Silicon Retinas (IOSIRE), developed for DVS data, is scalable for handling a wider application range. Adaptive M2M transmissions are favoured with layered data representation rather than local data processing, thus taking care of QoS, response times and energy constraints while sending only a selected amount of data to the cloud.

Current conventional IOT visual system deployments use low-end analytics and wireless networks, but the energy and bandwidth required by spatio-temporal resolution are infeasible for low-speed transmission apps. This is solved by compressive sensing and distributed video coding with protocols optimised for energy and information. But this isn't commercially available as yet and reconstructing the data from samples dynamically for DVC would be computationally infeasible for videos of high fps. Neuromorphic computing has emerged as a solution for DVS applications in the last few years, even though data encoding and transmission to servers remains unsolved. NC systems help estimate data traffic and in data sampling in constrained conditions. Up linking of M2M systems leads to layered data representation i.e. data is processed flexibly according to application, in addition to adaptivity and use of WSNs for collecting and processing compressed video frames. But these communication schemes/ protocols are not yet optimised for DVS-based apps.

Neuromorphic vision sensors use the address event representation protocol (AER) for asynchronously transmitting pixel-level intensity information. (x, y, t, p) is used, where x, y are pixel coordinates, t is the timestamp and p are the polarity i.e. increase or decrease in grey level intensity change. Motion of the neuromorphic sensor or change in illumination triggers events. Stationary vision sensors or static scenes don't transmit data, thus achieving the power and bandwidth constraints with low-latency. A diving scene example is used where green indicates increase in brightness and red indicates opposite. 8 data bytes are used to

represent this event, captured in 320X240 spatial resolution and an average neuromorphic event rate of 28.67 Keps (kilo events per second). The data rate of transmitting this video would then be about 1.83 Mbps, which is 10X lesser than the raw video transmission rate and with a higher temporal resolution. Knowing the data rate of neuromorphic visual sensors is important to decide apt transmission rates and as such, mean gradient approximations were used to measure video complexity. A linear relation is shared by sensor speed and event rate, and a 2-parameter exponential model is made to study relation between event rate and sensor speed and scene complexity. This results in 88.4% outdoor prediction accuracy by bits of model and an overall bit-rate accuracy of 84%, deeming it suitable for selection and design of the apt transmission strategy and performance evaluation of simulated neuromorphic-based IoT systems. Compression is required to overcome high data rates, and is innate in DVS during accumulation phase. Taking the event data's unique characteristics in interest, a lossless compression system is used, where a variable no. of bits (<8B used before) is used to represent neuromorphic events depending on dataset. Each studied event, albeit imaginary, is represented by 3.3 bits, but the compression rate in autonomous driving is much less (2.65 compared to 19.5). Out of all the listed strategies, Brotli method achieves the balance between speed and compression ratio. The challenges of complexity and delay limits remain.

Wireless communications systems are expected to handle data traffic with massive machine type communications (MTC) and ultra-reliable low latency communications (URLLC), while supporting spiking DVS and VR. Hence speed has to be cared for, keeping versatility in mind. To enable 5G end-to-end communication lasting <1ms, proactive computing and coded computing are some techniques used. Proactive computing means doing tasks before they're needed, and coded computing consists of eliminating unnecessary steps in processing tasks. Reducing the distance between sender and receiver is crucial for low-latency and is achieved by network densification and high-capacity data links (large data transfer channels). High data rates can be achieved by the mmWave band of the EM spectrum (30-300GHz), but lossy signals and high-power requirements are a challenge. Massive input massive output (MIMO), beamforming and other advanced signal processing techniques can counter the limits. ML methods are used to decide upon antennas and signal transmission, to reduce delay, based on factors like data queue, energy and communication channels' condition. A method used involves a framework inspired by partially observable Markov decision process, to optimise the time information takes for transmission while controlling energy usage. There is also a concept of multi-objective optimisation for minimising power use while maximising speed.

Object recognition has seen growth due to advances in DL, CMOS Active Pixel Sensing (CMOS APS). Neuromorphic Vision Sensing (NVS) is alternative to APS for energy and computationally efficient object recognition systems. Feature descriptors with classifiers like corner detectors, line/edge extractors, optical flow etc were used for researching, but low sales and high computational requirements are hurdles in their real-life implementation. Frame based methods of converting the neuromorphic events into synchronous frames of spike events, on which conventional CV techniques are applied is a method, however it doesn't capture the compact and async nature of NVS. The most common architecture is SNN utilising event-based method. Lack of suitable training methods stop SNNs from reaching the performance of gradient-based methods in learning complex relations. Since the activation functions are not differentiable, back-propagation is not an option, but researchers have devised a method of training NNs offline on continuous/rate-based models by

supervised learning and mapping the trained architecture on SNNs. But this yields low performance rates compared to gradient-based models. A solution to using NVS for object recognition is end-to-end graph-based framework, which uses graph CNNs and represents events as graphs to maintain event asynchronocity and sparsity, using traditional gradient-based back-propagation for training. On comparison, graph CNNs are fond to have lesser weights and 5 times less the ResNet. Hence, Graph based object recognition approach for NVS is a way to solve compact, spike-based, async nature of NVS with power of well-known GNN learning methods.

Recent research in spike-based compression, transmission and recognition in visual IOT systems is the paper's crux. The solutions described can be used for further developments in this field.

Name: URVI PATEL UID:22BIT034
Subject: IOT Date: 26/1/2024

A Review of Current Neuromorphic Approaches for Vision, Auditory, and Olfactory Sensors

Anup Vanarse, Adam Osseiran and Alexander Rassau

29th March, 2016

{Address-event representation, spatial contrast, temporal contrast, action potential, pulse-width modulation, Nyquist frequency, automatic gain control, resonant low pass filters, baseline sensor, drift behaviour, sensor-fusion}

As the data to be processed increased, Carver Mead introduced neuromorphic computing as a way to use the non-linear properties of transistors to conserve energy. Transistors were well-utilised as silicon retinas and cochlea to counter the high energy consumed by conventional methods. While several advancements have been made in the field of neuromorphic sensors, the authors bemoan the lack of a standard method for evaluating sensor outputs to gain benchmarks for improvement. They review some of the neuromorphic sensors innovated currently.

Neuromorphic Vision Sensors saw their start in silicon retinas in 1991. The currently used conventional frame-based approach generates redundant data. Reducing the capture rate results in information loss. Again, these methods consume much power, a problem that cannot be solved by even off-sensor processing. The rise of silicon retinas brought about energy-conservation and adaptive vision-sensing, while carrying out retinal functionalities, especially of cone cells. Taking the neurobiological model in consideration, an operational model could be made adhering to practical implementations. The difference in spatial and temporal contrasts is an important factor in solving this issue. With developments in AER, pixels can now report deviations in the contrasts, resulting in spiking outputs like the action potentials of ganglion cells. Thus, retinomorphic sensors use AER extensively. Adaptive photoreceptor circuits have played a role in developing Dynamic Vision Sensors (DVS). The DVS is highly effective in detecting changes in lighting much like our own eyes. Again, it works on AER and outputs spikes of pixel addresses where there was a change in luminosity. This is a significant benchmark for neuromorphic vision sensing. An alternative to DVS is frame-based temporal detection imagers, that compute difference in photocurrent of each frame. However, they are slower in responding and range. The DVS gives faster response, has a dynamic range, sub-millisecond precision and consumes low power, making it useful in robotic solutions and real-time systems. The capabilities of DVS in terms of output are tripled by addition of PWM. A hybrid of frame-based and frame-free model is adapted as DAVIS (Dynamic and Active-pixel Vision Sensor). This is a mix of async, independently functioning DVS sensors and synchronous active-pixel vision sensors like the ones used in traditional cameras, and hence is frame-free and frame-based. This helps create a benchmark for research in static spatial information-processing and detecting changes with respect to time (dynamic temporal changes) with minimal latency.

Auditory sensors use sampling of data at a specific Nyquist frequency for sensing, leading to power-hogging analogue-to-digital conversions (ADC). Reduction of sampling may save power, but lead to information-loss. The inception of bionic ears started with making of a VLSI-based cochlea, having features like automatic gain control, resonant low-pass filters,

delay. Initially lacking in biasing circuits, issues like device mismatch and dynamic range were improved in further iterations, leading to a benchmark- Overlapping cochlear cascades with a 61dB and 0.5mW range. This led to the development of battery-powered silicon cochlea with similar characteristics of adjusting amplitude according to distance and pre-amplifying signals. Such processors, being energy-efficient, can be used in low-power speech-recognition front ends. Auditory scene analysis too is a vital aspect, and can be implemented by AER. An AER EAR is developed using a silicon cochlea and MEMS mic (Micro-Electrical-Mechanical-Systems). Here 32 filters and channels are used to analyse different sound aspects to better evaluate its source. An advanced version AER EAR2 is developed which uses 64 channels and pre-amplifiers with individual channel adjustments. This is useful in localised applications like recognising speakers. Because it uses a spike-based approach, it is 40 times more efficient than conventional methods. The information can be precisely captured and transmitted via USB, allowing for speaker identification. Applying innovative techniques like spike-based audio front-end allows for more research in this field.

Mechanical concepts are used to determine and measure odours. However, conventional electrical noses have high manufacturing costs and are not portable, in spite of high reliability in detecting target gases. Using CMOS and MEMS combined with advanced pattern matching leads to new sensing methods. Bio-based olfaction units have sensor-array, signal-condition circuitry and a pattern-recognition unit. Neuromorphic computing helps combine all these onto a single chip with neural network implementation. The variations in operating current and sensor poisoning can be nullified by use of CB polymer sensor array on AMS 0.6 micro-meter CMOS process. This proved to be a technological benchmark, as the neuromorphic implementation helped detect the target gas faster even with background odours while utilising less power due to the implementation of STDP (Spiking Time Dependant Plasticity). A bio-inspired mucosa too was developed in 2007 but it had an unsatisfactory response-time. Researchers have realised it is better to emulate only the critical parts rather than entire olfaction pathway. A Tin-Oxide Gas sensor array was developed in 2011.this had rows of sensors having similar drift behaviour. They use same catalyst for each group to detect a wide range of chemical gases. the firing delays in spiking output helps generate a drift-insensitive signature response for specific gases, which are stored in 2D array of spatio-temporal spike patterns. Implementing single CMOS chip helped in low power-consumption (6.6mW) and high identification accuracy (94.9%). An E-Nose using polymer sensors and interface circuitry with ADC, employs the signature-spiking-response technology described above. However, its pattern-matching technology proves to be computationally costly. This is commercially used in Cyranose 320 and has low-average-power-consumption (3.6 micro-watt) and moderate testing accuracy (87.59%). The STDP Learning Rule, used for detecting fruity odours, uses sensors from Cyranose 320. It can identify 3 odours concurrently and has the same average power consumption and testing accuracy as E-Nose. The NEUROCHEM project works to make a large polymer sensor array which are conductive, redundant and sensitive to certain odours, much like biological Odour Receptor Neurons (ORNs). Neuromorphic Olfaction requires improvements in response time, interfacing, pattern-matching, and signal-conditioning.

The authors specify trends in sensor-fusion architecture by stating the dearth of spike-based data format despite benefits of low-power consumption and sparse output-data-generation. Conventional frame-based or audio-signal processing methods thrive due to years of research

whereas neuromorphic computing uses event-based processing, which is still in nascent stage. Advanced research has led some neuromorphic prototypes to be developed into commercial products like DVS128PAER and DAS1. The Koala robot for object tracking uses sensor fusion of Audio-Visual Neuromorphic Sensors. CAVIAR, BrainScaleS, and SpiNNaker are some projects which promote sensor fusion with data correlation. These systems have numerous uses in robotics and environmental monitoring.

The authors of this paper have thus studied some of the research in neuromorphic approaches for visual, auditory and olfactory sensors. The development of AER, DVS and DAVIS plays a crucial role in both vision and auditory sensing. However, more research is required in the field of olfactory, in terms of evaluating models and benchmarks. The authors suggest the need for correlating different sensor inputs and further research for vibration and pressure sensors.

Name: URVI PATEL UID:22BIT034
Subject: IOT Date: 26/1/2024

Powering Next-generation Industry 4.0 by a Self-learning and Low-power Neuromorphic System

Hongyu An, Dong Sam Ha, Yang (Cindy) Yi

7th October 2020

{associative learning, synaptic connecting strength, distributed data learning, reservoir computing}

As we approach the 4th industrial revolution, the last of which brought upon the onset of digital communications and integrated circuits, we see that humans are still not freed from the tediousness of the product line. What we require in the 4th wave would be an autonomous, self-learning manufacturing system that decides for itself. The requirement of smart sensors, sophisticated communication and the innate ability to control processes within and without can be covered by Cyber-Physical Systems (CPS), IOT, and next-level AI. These would help in mapping the real-world into cyberspace, empowering sensors to collaborate and provide the autonomous nature through cognitive, cloud and AI computing. Current AI systems don't hold a torch to the low-power, self-learning system expected. ANNs, albeit powerful, are limited in efficiency and performance by the von Neumann structure, whereas human brains perform a superior task of pattern recognition, reasoning, control, movement within a mere 20W. Industry 4.0's AI must aim to imitate the human-brain through Neuromorphic Computing. Specifically, the goal is to make an organ-like sensory system to receive and convert real-world signals into spiking ones for IOT, and to develop a human-like learning system, so that the neuromorphic system learns by its own experience. Data exchange through wireless IOT will play a crucial role in achieving this aim. However, data analytics is among the many fields that are challenging to deploy through IOT. Problems faced due to power can easily be solved by ANNs (high power can be consumed) or Neuromorphic Computing (economical in power).

Due to powerful computational resources on the server-side, machines in a factory can communicate with each-other, access resources easily and collaborate on tasks. The von Neumann architecture limits the abilities of these machines by bottlenecking the data communication due to CPU and memory being at different locations. The task of recognising 1000 objects, taking 250W in today's machines, can be achieved by our human brain at 20W, and at a faster pace. The intrinsic structure of the human brain, with its millions of synapses forming a neural network is capable of associative learning and spike-based signal representation. The low firing rate of spikes helps the brain operate in a highly energy-efficient manner. Adding to the DL abilities presented by neural architecture, the associative memory helps brain learn from surroundings. A network heavily inspired by this is proposed in the paper for industrial use. Moreover, the network-based neuron system is not only existent in the brain, but throughout the entire body, capturing external signals on low frequency, therefore consuming less energy. The iniVation Dynamic Vision Sensor is one such example of a biologically-inspired neuromorphic commercial system. These are trained to be adaptive and reactive to the real-time changes, making it suitable for complex manufacturing tasks in Industry 4.0. The neural-networks based robots would help in designing advanced robotics for autonomous manufacturing.

Instead of using Boolean-value-based von-Neumann-architecture, neural-network based architecture is proposed, mainly, Distributive, Cluster, Associative Neuromorphic Computing Architecture (DNCA, CNCA, ANCA). In DNCA, neurons (computing units) and synapse (memory units) are placed so that distance between them is minimised, thus reducing energy spent in signal propagation during computation. CNCA divides DNCA into multiple sub-regions for the different senses to achieve the parallel computational capabilities of the brain. The outputs of CNCA are correlated for system to learn from surroundings and experience by ANCA.

The increasingly large dataset sizes remain a hurdle for deploying ANNs, especially since the GPU size hasn't increased as much. The tediousness of the task in making datasets leads to the rise of associative-learning, wherein the system learns the relationship between 2 concurrent events. This leads to a decrease in need of large datasets. The 2 factors weighing in associative learning are: synaptic connecting strength modification and distributed data processing. The connection between sensory and response neurons increases when learning occurs. Images and audio, being 2 different types of data are processed in 2 different locations of the brain. The associative learning is done between auditory and visual data by studying the simultaneous probabilistic output scores of ANNs for both modes. Such a neuromorphic system is important in speeding up the learning process essential for an autonomous manufacturing system in Industry 4.0.

Reservoir computing, a neuromorphic algorithm belonging to the category of RNN is a beneficial aspect for Industry 4.0. RNNs are not defined by current state only but by prior states too, similar to a biological neural network. RNNs can be computationally expensive due to the training of all weights within network. RC solves this problem as it employs a self-connecting model similar to biological neurons, hence making it suitable for employment in character and speech recognition. RC reduces the computational complexity in low-power consumption, making it suitable for use in Industry 4.0. the computational accuracy of RC depends on number of neurons in its reservoir layer. Adding time-delay to the mix makes these systems highly hardware-friendly and richly dynamic. Using only 1 non-linear neuron in every reservoir, stacking these reservoirs in a non-linear manner helps achieve computational efficiency and accuracy. These make RC a good candidate for mobile, low-power devices in Industry 4.0.

ANNs in IOT can be used to study data patterns in sensors. They can also use reinforcement learning to dynamically adjust bandwidths. Identification and classification of data helps in filtering the data used for analytics. User behaviour can also be predicted to perform operations in advance. The energy and computational requirements of ANNs in IOT remain a challenge, as do the use of apt data and protocols for formatting, training and security. The author gives the example of cybersecurity to demonstrate how ANNs may be used for a good purpose (threat detection) and a bad one (intrusions). WSNs can be used to achieve a low-power communication system in IOT ANNs, to perform simultaneous data processing and transfer. Out of 8 ML models studied in a cited paper, Deep Learning ANNs (DLANN) have the best classification power, but the longest execution time. The use of Laguerre neural network- based dynamic programming to improve tracking mechanisms, CNNs for image detection, and ANNs for target surveillance is studied. These demonstrate what are the possible use cases of neural networks in IOT. The paper emphasises that ANNs are without

doubt an important tool for solving a variety of problems in IoT like communication quality, classification of activities, tracking efficiency, intelligent data analytics, and smart operation.

The 4th stage of industrial revolution brings forth low-power, autonomous systems, the capabilities of which can't be fully covered by von Neumann architectures and current deep learning trends. Neuromorphic computing, with its aim to emulate the brain provides a welcome change in low-energy consumption and self-learning. This paper covers the emerging architectures, associative memory learning and reservoir computing. It also delves in the use of ANNs in IOT.

A Neuromorphic Approach to Image Processing and Machine Vision

Arvind Subramaniam

21-23 December 2017

{synchronised oscillation, Random Walker Algorithm, anisotropic diffusion, subthreshold domain, async communication, Asynchronous Time-based Image Sensor, Top-down approach, Bottom-up approach, volitional selection, systolic array, Boltzmann Machine models, combinatorial optimisation problems}

Neuromorphic Computing is a young, actively researched field, much like artificial vision. Despite years of effort, getting the computer to match the biological vision system is a challenge. Moreover, implementation of computer-vision is a power-intensive task and often unreliable. Neuromorphic computing allows us to perform computation and memory parallelly, to save energy and mimic biology, in autonomous audio and visual processing systems. Image segmentation, the study of dividing the image into parts to better glean relevant information out of the process, can be done through neuromorphic circuits. Photo-detector cells receiving similar light intensity oscillate together, making it synchronised oscillation. This has led to the development of Neuromorphic Vision Sensors, which has further led to the use of attention mechanism and object recognition.

Image segmentation algorithms may be supervised, unsupervised or semi-supervised. The Random Walker Algorithm, a semi-supervised one, treats every pixel as a node having weighted edges between its neighbouring pixels. These edge weights, grouped together by similarity of pixels, sum up to the electrical conductance. The algorithm labels few initial pixels as seeds to initiate pixels, and also to complete segmentation when every node has been assigned a seed to classify it. It would be computationally intensive to perform a partial differential equation on the seeds, which is why these graphs are viewed as electrical circuits. Each node at foreground is assigned an electric potential- whose value will be equal to probability that walker reaches background node before foreground node. Manually changing the resistance for this task makes it unfeasible to realise in conventional CMOS. An alternative approach is Anisotropic diffusion which removes noise from images by smoothening out pixels from only one side of the image. This is unlike isotropic diffusion, where the average of all edge pixels is taken, leading to unnecessary smoothening and loss of soft edges. To implement this, a memristor-based crossbar array is used. This changes resistance between node and a foreground/background seed based on history of current flowing through edge. The memristor's variable resistance changes edge weights when there is change in potential between nodes after every iteration, instead of doing so manually.

Synchronised oscillations help in image segmentation, since it helps distinguish objects through sets of oscillating detector cells. An example is locally excitatory globally inhibitory oscillator network (LEGION). However, it is sensitive to noise, leading to fragmentation. However, by suppressing oscillators corresponding to noisy parts of image may lessen the problem. A memristor-transistor pair can be used to design such a non-linear oscillatory system. the transistor helps counter the non-linear relation of memristive resistance with time.

Neuromorphic vision sensors reduce the computational load and offer visual perception, making them popular in the field of artificial vision. The making of retina with analogue

electrical circuits helps solve 2 or more visual problems simultaneously, consuming less power. This is because of the circuits being parallel and operating in sub-threshold domain. This allows us to use computationally dense detectors for async communication since synchronous communication has the problem of misinterpretation. The concept of using silicon retinas and silicon cochlea has been introduced due to address event representation (AER) in the 1990's. this helps in async transmission of information. In AER, a variable number of lines are used to communicate the data, which is an address of an analog element present on sender. The synchronisation of communication is enabled by ACK and REQ which are active low lines. The power consumed is reduced as signals are transmitted according to the activity of each individual pixel. Asynchronous devices responding to changes in pixel brightness (DVS) reduce data storage and computational complexity. STDP can be used to extract correlated spatio-temporal features collected by DVS. These make use of memristive technology used in Asynchronous Time-based Image Sensor (ATIS) that also measures conditional exposure.

Visual Attention is used in our biological eyes to focus on what is important and not pay attention to the redundant details. This must be mimicked in machines too, for processing only the important regions of the image. This attention mechanism is of 2 types:

1. Goal-oriented Top-Down (TD) approach:
   This is known as sustained/endogenous attention, and directs attention to certain features and images in space for a sustained amount of time, making it slower than an involuntary, bottom-up approach. It's suitable for tasks that require special concentration, like driving.
2. Image-Driven Bottom-Up (BU) approach:
   This is meant to capture unexpected events that may stand out, like accidents on-road. These may be captured by vision involuntarily. The bottom-up approach takes spatial discontinuity in colour, orientation and creates a topological saliency map, which is scanned in order of decreasing saliency.

Further on, there are 2 types of TD approaches too: volitional TD selection process and mandatory TD selection process. A volitional selection means the person wilfully selects what to focus on. Mandatory selection is the opposite, especially in case of optical illusions where person compulsorily focuses on the vase and not on two faces, while volitional switches flexibly between two. Integrating both TD and BU is essential for use-cases such as surveillance systems, to detect unusual events. TD approach helps use learnt knowledge to increase the weightage of focus in certain areas, while BU approach gives increases saliency to the target.

Visual object recognition, a computationally complex problem, refers to perceiving and identifying the physical properties of an object. The Hierarchical Model and X (HMAX) is a model commonly used for this purpose. The inner workings of this model are described in detail. The objective of this model is to obtain feature vectors from grey-scale images and to classify these vectors using an SVM. Implementing HMAX is difficult on low-power devices due to excessive energy-consumption required by it. The use of hardware accelerators, like FPGA and ASIC help solve this problem. FPGA provides best functional configurability and ASIC provides high efficiency. Recent studies employed a memristive Neuromorphic Computing Accelerator (NCA). This in turn uses crossbar arrays, formed by intersecting

memristors, as opposed to conventional accelerators based on systolic (parallel) computing arrays. Boltzmann Machine models, designed with memristors enhance neural computation efficiency and energy efficiency in solving combinatorial optimisation problems (finding best solution out of a finite set of a number of solutions). In comparison to a standard RRAM-based memory, the memristive hardware accelerator demonstrates a significant improvement, achieving a 6.89x boost in performance and a 5.2x reduction in power consumption. However, the proposed memristor-based accelerator has not been shown to solve problems of higher computational complexity yet.

This paper expands upon neuromorphic techniques used to optimise the algorithms employed for solving critical vision problems. These include the use of memristive crossbar arrays as edges between nodes and foreground/background seed. The mechanisms and types of neuromorphic vision sensors too have been explored the application of neuromorphic memristors, non-volatile memory devices and CMOS has resulted in achieving vision tasks with energy-efficiency. The author hopes that the utilisation of neuromorphic technology increases in IOT and quantum computing in the future.

Name: URVI PATEL UID:22BIT034
Subject: IOT Date: 26/1/2024

Conclusion

Neuromorphic Computing is the fresh new technology- aiming to combine deep learning and electronics in the race to achieve energy-efficiency. This proves it to be suitable for constrained devices while mimicking the human brain. Research has shown its extensive applications. Due to my interest in Computer Vision, I have focused on studying about Dynamic Vision Sensors and asynchronous, event-driven data processing. This is much like what happens in our brain, and can help in building autonomous manufacturing and navigation systems. As we march into a future of chips being installed in human-brains, neuromorphic computing has a scope for greater research and use-cases. Maybe we can really eliminate a lot of the problems that intimidate us, like paralysis, blindness, through brain-inspired neural networks in IoT. Maybe, Musk’s Neuralink is just a start.

References

1. Energy and Area Efficiency in Neuromorphic Computing for Resource Constrained Devices- Gangotree Chakma, Nicholas D. Skuda, Catherine D. Schuman, James S. Plank, Mark E. Dean, Garrett S. Rose (23rd May, 2018)
2. The Human Brain Project and Neuromorphic Computing- Calimera, Enrico Macii, Massimo Poncino (17th Oct, 2013)
3. Challenges and Perspectives in Neuromorphic-based Visual IoT Systems and Networks- Maria Martini and Nabeel Khan, Yin Bi and Yiannis Andreopoulos, Hadi Saki and M. Shikh-Bahaei (09 April 2020)
4. A Review of Current Neuromorphic Approaches for Vision, Auditory, and Olfactory Sensors- Anup Vanarse, Adam Osseiran and Alexander Rassau (29th March, 2016)
5. Powering Next-generation Industry 4.0 by a Self-learning and Low-power Neuromorphic System- Hongyu An, Dong Sam Ha, Yang (Cindy) Yi (7th October 2020)
6. A Neuromorphic Approach to Image Processing and Machine Vision- Arvind Subramaniam (21-23 December 2017)