Applicant: Thomas Shoesmith

Supervisor: Thomas Nowotny

Mentor: James Knight

Non-Technical Abstract

Modern Artificial Neural Networks (ANN) still require a lot of processing power. For instance, AlphaGo used approximately 1MW of power to run, whilst a human brain only uses 20W.

Unlike units in ANNs which continuously exchange real-valued activations, neurons in spiking neural networks communicate infrequently using binary events called spikes like neurons in biological brains. This reduced communication means that, theoretically, SNNs could offer a much more energy-efficient alternative to ANNs.

Training SNNs from scratch remains tricky but there are several promising approaches for converting already trained ANNs to SNNs. The current limitation with SNNs is that they require a lot of signal spikes in order to achieve the same task performance as ANNs; the aim for this project will be to investigate more efficient means of spike based encoding of activations in SNNs in order to improve on both, the number of signals required as well as the accuracy of the SNN in image classification.

Technical Abstract

Despite yielding impressive results from image classification to playing against humans in games such as AlphaGo [Silver et al., 2016], Artificial Neural Networks (ANNS) require a vast amount of energy when compared to the brain. AlphaGo as an example required approximately 1MW of power to run [Mattheij, 2016], meanwhile the energy consumed by an average human brain is equivalent to approximately 20W [Ling, 2001] giving AlphaGo 50,000 times more energy in this competition. This brings us onto the latest challenge in modern machine learning which is the focus on the efficiency of algorithms so that they can be run on more portable devices or scaled up without the requirement of large energy resources.

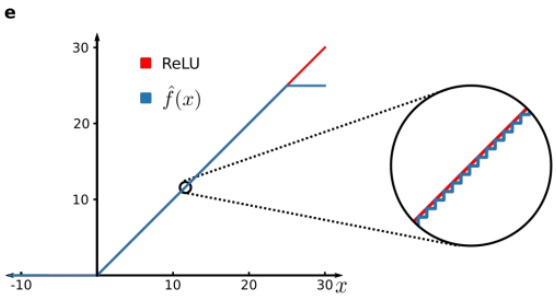
Spiked Neural Networks (SNN) based have shown the potential to be more energy efficient than comparable ANNs mainly due to their lower computational requirements and the sparser spike-based communications between neurons [Khacef et al., 2018]. However, training large SNNs from scratch remains challenging and converting a high performing ANN into a similarly performing SNN has typically involved encoding ANN activations in the firing rate of the spiking neurons [Sengupta et al., 2019]. However, the resulting SNN required a relatively large number of spikes to generate similar performance and these larger numbers of spikes outweighed the energy savings of the SNN.

In a recent paper [Stöckl and Maass, 2021], the authors introduce a new method for converting ANNs to SNNs using the so-called “Few Spikes (FS) conversion”. FS treats each neuron’s output spikes as a binary code representing the activation of the ANN neuron and, by using the timing of the spikes to hold additional information in this way, fewer spikes are needed and each input can be presented for fewer timesteps.

This project will build on the research from this paper to explore further optimisations to these binary codes allowing the number of spikes and timesteps to be further reduced and efficiency improved. The SNN performance will be evaluated by its accuracy of image classification on a selection of benchmark data sets including CIFAR-10 and ImageNet. CIFAR-10 is a database consisting of 60,000 32x32 colour images that can be classified into 10 classes with 6000 images per class. ImageNet is a much more challenging dataset consisting of around 14 million larger images in 20000 classes.

Primary Objectives

1. Implement a FS neuron using the ReLU Activation Function (AF) [f(x) is the AF from the paper by Stöckl and Maass, 2021]



2. Experiment with smarter, more efficient codes for encoding the ReLU AF instead of the simple binary encoding of [Stöckl and Maass, 2021] (e.g. Huffman codes)

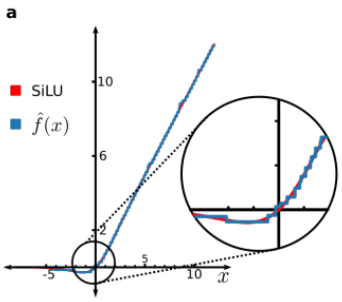
3. Experiment with AF input value representation
Current methods for AF encoding in essence use base 2 fixed-point arithmetic. This results in an inability to represent smaller numbers, in particular when using only a few spikes/time steps (bits). Being able to store values in the equivalent of floating point would resolve this allowing for a better representation of inputs and AFs with fewer spikes

4. Work with mentors to implement these improved models in mlGeNN (https://github.com/genn-team/ml_genn/tree/master/ml_genn) - a library developed at the University of Sussex for SNN-based ML.

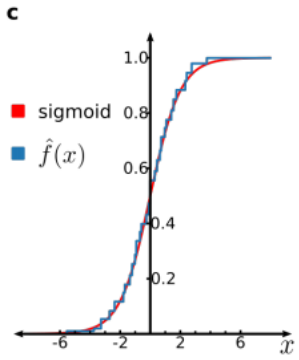
5. Evaluate performance of the new encodings on CIFAR-10.

Extended Objectives

1. Experiment with activation functions of SiLU



2. Experiment with activation functions of Sigmoid



3. Evaluate performance of new models using ImageNet

Relevance

I am extremely interested in the field of machine learning and this project is inspired by discussions with the supervisor and influenced by lectures from the modules taken this term. With such a big focus on machine learning in computer science today, it is important to be able to write algorithms as efficiently as possible for time saving, energy savings and accuracy.

This project is extremely relevant to the fields of Machine Learning (ML) and Neural Networks (NN) within Computer Science. Being able to reduce the energy demands of NN will allow for inference to be moved from the data centre to the edge improving the autonomy of portable devices such as: voice assistance (Apple's Sir and Amazon's Alexa to name a couple). Additionally, more efficient inference could allow medical facilities to analyse fMRI images more quickly without using remote services to handle the heavy processing. This is not only desirable for better efficiency but also solves many issues around privacy and security. In the long term the proposed research is aiming towards better sustainability of future AI systems in the context of novel, neuromorphic computing systems.

References

Stöckl, C. and Maass, W., 2021. Optimized spiking neurons can classify images with high accuracy through temporal coding with two spikes. *Nature Machine Intelligence*, 3(3), pp.230-238.

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. and Hassabis, D., 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), pp.484-489.

Mattheij, J., 2016. *Another Way Of Looking At Lee Sedol vs AlphaGo · Jacques Mattheij*. [online] Jacquesmattheij.com. Available at: <https://jacquesmattheij.com/another-way-of-looking-at-lee-sedol-vs-alphago/>

Ling, J. (2001). Power of a Human Brain.
<https://hypertextbook.com/facts/2001/ JacquelineLing.shtml>

L. Khacef, N. Abderrahmane and B. Miramond, "Confronting machine-learning with neuroscience for neuromorphic architectures design," 2018 International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1-8, doi: 10.1109/IJCNN.2018.8489241.

Sengupta, A., Ye, Y., Wang, R., Liu, C. and Roy, K., 2019. Going Deeper in Spiking Neural Networks: VGG and Residual Architectures. *Frontiers in Neuroscience*, 13.

Nrdc.org. 2021. [online] Available at: <https://www.nrdc.org/sites/default/files/gadget_report_r_19-07-b_13_locked.pdf>

Apple Machine Learning Research. 2021. *Hey Siri: An On-device DNN-powered Voice Trigger for Apple's Personal Assistant*. [online] Available at: <https://machinelearning.apple.com/research/hey-siri>

Cs.toronto.edu. 2021. *CIFAR-10 and CIFAR-100 datasets*. [online] Available at: <https://www.cs.toronto.edu/~kriz/cifar.html>

ImageNet 2021. [online] Available at: <https://imagenet.stanford.edu/index.php>