Impact of time discretization on the efficiency of continuous time Spiking Neural Networks

TU Delft

The effects of the time step size on the accuracy, sparsity and latency of the SNN.

1. Spiking Neural Networks

- Recent advancements in deep learning, such as Large Language Models, come with high computational costs. For instance, training GPT-3 consumes around 305% of the carbon emissions of a full passenger flight from New York to San Francisco [1]. This is why research is interested in more power-efficient alternatives.
- Spiking Neural Networks (SNNs) offer a more energy-efficient alternative to Deep Neural Networks by mimicking the brain's natural processing using events called spikes. Although SNNs may lag in accuracy compared to deep neural networks, they excel in energy efficiency.
- These models can be even more energy efficient when implemented on digital neuromorphic chips such as Intel Loihi and IBM TrueNorth. These chips are designed to handle the unique processing needs of SNNs like sparse and event-driven computations [3].

2. Challenges

- A lot of the state-of-the-art performance of SNNs in recent research has been achieved through supervised learning models that leverage intricate error backpropagation techniques in continuous time.
- This imposes a new challenge when converting these mechanisms on a neuromorphic chip. Because time is discrete on digital hardware numerical errors can be introduced as we can not calculate the infinitely precise value of variables depending on time.

3. BATS Model

- In BATS [2] a CUBA (Current-Based) LIF (Leaky-Integrate-and-Fire) neuron is used with a soft reset of the membrane potential. Soft reset implies that when the membrane potential of the neuron passes a threshold value it is decreased by a leak value and effects are propagated to neurons in the next layers.

4. Accuracy, Sparsity and Latency

- The main reason spiking neural networks excel in efficiency is their sparsity. Even if networks can have many neurons computations happen only when the neurons are activated.
- We aim to optimize spiking neural networks for power efficiency without significantly compromising their accuracy, ensuring they remain competitive with traditional artificial neural networks.
- On top of accuracy, it is also desired to have networks with low prediction latency. Being able to obtain the same amount of information with earlier spikes increases the idle time of the network.

5. Time Discretization

When doing time discretization the spike at time t will be delayed to the end of the time bin it is situated in. This method of doing time discretization allows us to perform experiments without having high numerical errors or affecting the network drastically.

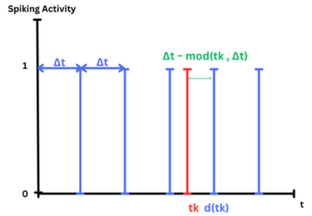

Figure 1 : Visualisation of the time discretization

6. Straight-Through Estimator

The time discretization function is non-derivable with respect to the spike timing. Therefore we apply a naive method to be able to backpropagate effectively in the SNN. More specifically we approximate the loss derivative with respect to the continuous spike timing to the loss derivative with respect to the discrete spike timing.

7. Experiments and Environment

The main datasets used will be the MNIST, EMNIST and Fashion MNIST datasets. We run experiments to measure the sparsity and prediction latency of the model after adapting it to discrete time using different time step sizes.

- **Sparsity** will be measured based on the spike count of neurons after training.
- **Latency** will be measured using the time (ms) to reach a certain confidence level during training.
- **Accuracy** will be measured only on the test set.

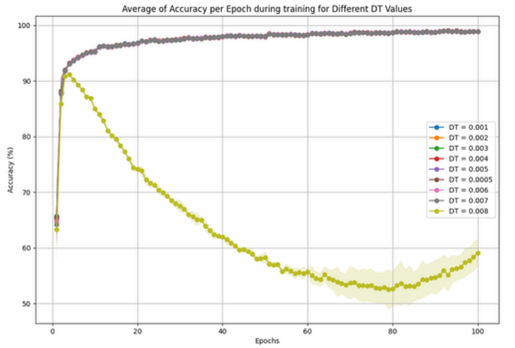

Figure 2 : Average Accuracy% on MNIST dataset during training

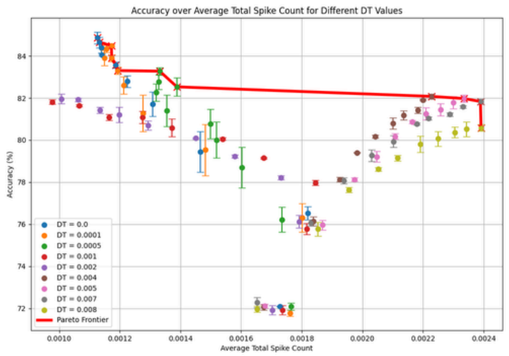

Figure 3 : Accuracy % on the test set for EMNIST dataset over the inverse of the average total(hidden and output layers) spike count. The red line crosses the points on the Pareto Frontier.

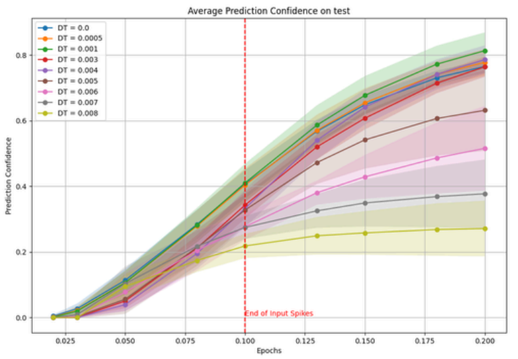

Figure 4 : Prediction confidence over time for different Δt on the Fashion MNIST dataset

8. Conclusions and Future Work

- With this research, we have shown that physical limitations of time on hardware devices do not present drastic errors in the accuracy of the Spiking Neural Network. On the other hand, setting a suitable timestep can have beneficial effects such as improving sparsity while maintaining a high accuracy. However, our analysis is still limited by errors introduced from the backpropagation approach.
- Future work could include an analysis of the errors introduced by changing the backpropagation methods.
- Another recommendation for future work could be using a surrogate gradient approach to better approximate the derivative of the discrete spike times.

9. References

[1] - D. Patterson, J. Gonzalez, Q. V. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "Carbon emissions and large neural network training," arXiv preprint arXiv:2104.10350,2021

[2] - F. Bacho and D. Chu, "Exploring tradeoffs in spiking neural networks," Neural Computation, vol. 35, no. 10, pp. 1627–1656, 2023.

[3] - K. Eshraghian, M. Ward, E. Neftci, X. Wang, G. Lenz, G. Dwivedi, M. Bennamoun, D. S. Jeong, and W. D. Lu, "Training spiking neural networks using lessons from deep learning," Proceedings of the IEEE, vol. 111, no. 9, pp. 1016–1054, 2023.

Alexandru Gabriel Cojocaru | Supervised by: Nergis Tömen, Aurora Micheli