ANN-to-SNN Conversion: Enabling Energy-Efficient Machine Learning for Edge Devices

Md Asiful Hoque Prodhan

The University of Texas at San Antonio, San Antonio, TX 78249

Abstract

Goal: Improve energy efficiency of machine learning on edge devices using Spiking Neural Networks (SNNs).

Why SNNs: Brain-inspired networks that process information via discrete spikes, enabling significant power savings over traditional deep neural networks.

Approach:

- ❑ Train an Artificial Neural Network (ANN) with standard or customized activation functions.
- ❑ Convert the trained ANN into an SNN for event-driven, low-power computation.

Benefit: Combines the high accuracy and training efficiency of ANNs with the energy efficiency of SNNs.

Impact: Supports real-time, low-power AI applications in areas such as IoT devices, embedded systems, smartphones, and autonomous systems.

Introduction

SNNs for Efficiency: Spiking Neural Networks offer low-latency, energy-efficient inference, especially when deployed on neuromorphic hardware.

Key Difference: Unlike Artificial Neural Networks that use continuous activation functions, SNNs transmit information via discrete spikes.

Biological Plausibility: This spiking mechanism provides a computational model closer to how the brain processes information.

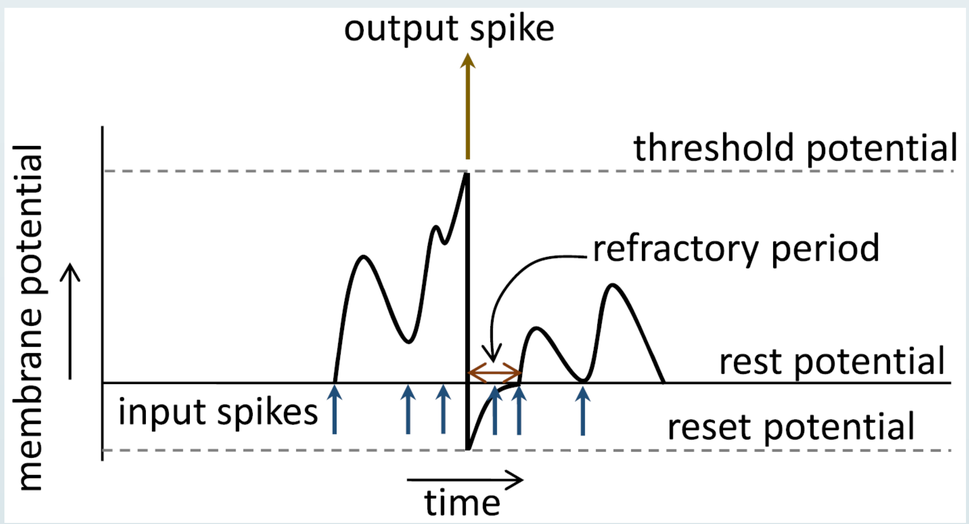

Figure 1. Membrane potential dynamics during spiking activity [1].

Main Challenges:

- ➢ Training SNNs is difficult due to the non-differentiable nature of spike events.
- ➢ Mapping continuous ANN activations to discrete spike events remains a key technical hurdle.

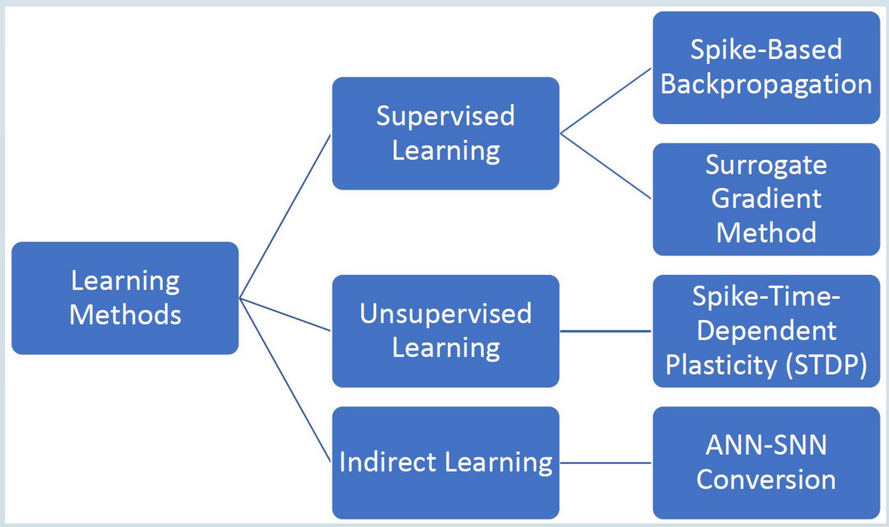

Figure 2. Learning Methods in Spiking Neural Networks .

Methodology

Approach 1 – Indirect ANN-to-SNN Conversion

- Train an Artificial Neural Network (ANN) using conventional methods.
- Convert the trained ANN to an SNN using the Sinabs library [2].
- Leverage ANN training efficiency and SNN energy efficiency.
- **Target:** Maintain ANN accuracy while enabling low-energy, spike-based computation.

Approach 2 – Spike-Compatible Training (Inspired by [3])

- Train ANN architectures (VGG-8, VGG-16, ResNet-18, ResNet-20) from scratch with custom spike-compatible activation.
- Activation output quantized to match Integrate-and-Fire neuron behavior.
- Convert to equivalent SNN without re-training, transferring activation parameters.
- **Goal:** Achieve low-latency, high-accuracy SNN inference with few time steps.

Results

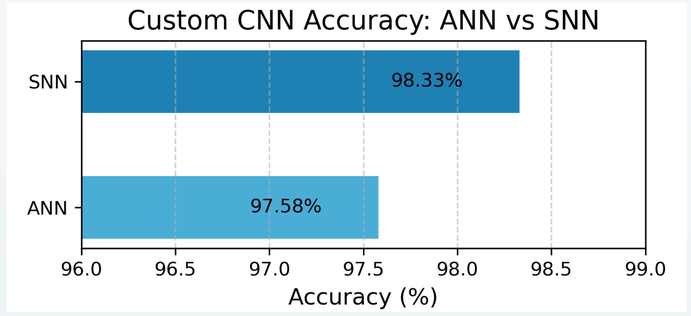

Figure 3. Custom CNN accuracy on MNIST dataset for ANN vs. indirectly converted SNN.

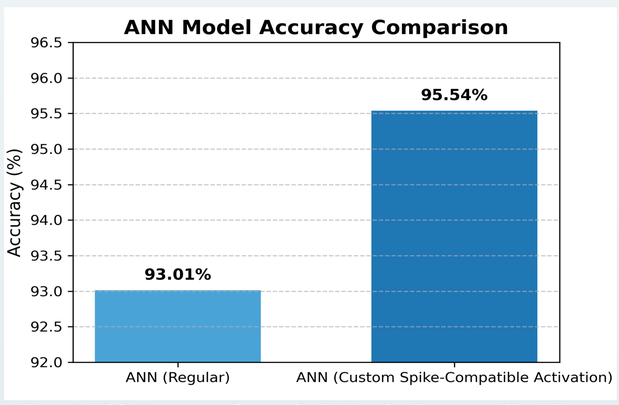

Figure 4. VGG-16 accuracy for regular ANN and spike-compatible ANN on CIFAR-10 dataset.

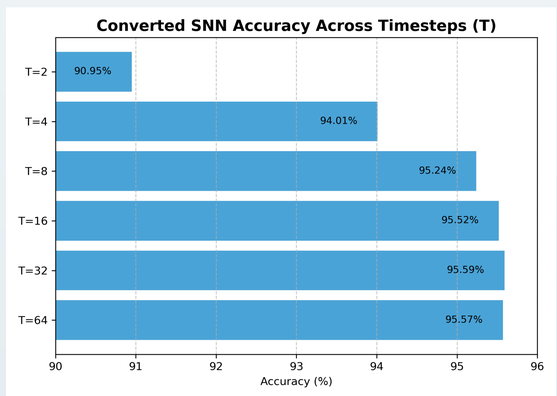

Figure 5. VGG-16 accuracy for the converted SNN on CIFAR-10 dataset.

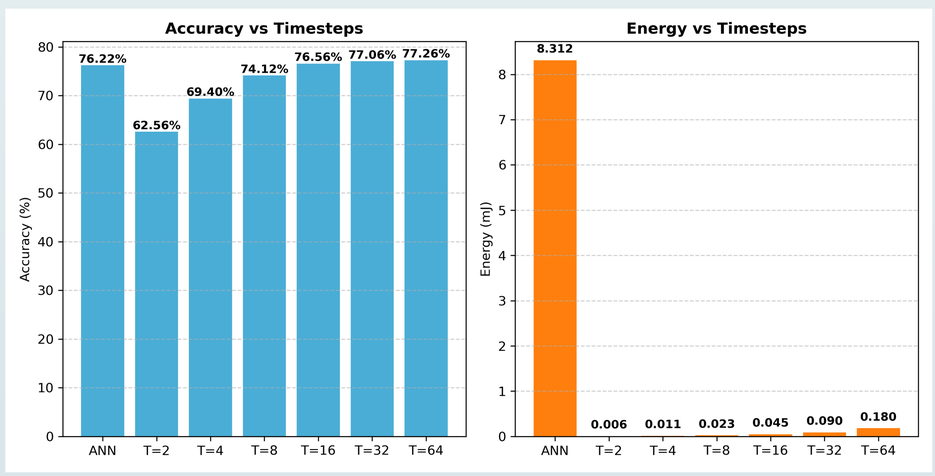

Figure 6. Accuracy (left) and energy consumption (right) vs. timesteps for VGG-16 on CIFAR-100 dataset.

Conclusion

- Both investigated approaches confirm that SNNs can achieve comparable accuracy to ANNs while offering potential for improved energy efficiency.
- Indirect ANN-to-SNN conversion on MNIST retained performance with minimal accuracy drop after conversion.
- Spike-compatible training on CIFAR-10 achieved high accuracy SNN inference with a small number of simulation time steps, enabling low-latency operation.
- Ongoing evaluations aim to further validate spike-compatible training for large-scale datasets and architectures, targeting latency and energy savings for deployment in resource-constrained environments.
- Future work will explore tuning conversion parameters, optimizing spike-compatible activations, and testing on real neuromorphic hardware.

References

[1] N. Rathi, I. Chakraborty, A. Kosta, A. Sengupta, A. Ankit, P. Panda, and K. Roy, "Exploring Neuromorphic Computing Based on Spiking Neural Networks: Algorithms to Hardware," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–49, 2023, doi: 10.1145/3571155.

[2] "Sinabs Documentation," https://sinabs.readthedocs.io/v3.0.2/index.html (accessed Aug. 1, 2025).

[3] T. Bu, W. Fang, J. Ding, P. Dai, Z. Yu, and T. Huang, "Optimal ANN-SNN Conversion for High-Accuracy and Ultra-Low-Latency Spiking Neural Networks," arXiv preprint arXiv:2303.04347, 2023.