DynSNN: A Dynamic Approach to Reduce Redundancy in Spiking Neural Networks

Fangxin Liu[1], Wenbo Zhao[1],Yongbiao Chen[1], Zongwu Wang[1] and Fei Dai[2]
1. Shanghai Jiao Tong University; 2. Academy of Military Sciences

icassp 2022
Singapore, China, Online

SNN PRELIMINARIES

- **ANN:** High Precision, Dense Computation.
- **SNN:** Low Computation cost.
  - **Directly Trained SNN** (Low Precision, Moderate Delay)**:** utilizes the spike-based BP algorithm or STDP-based learning algorithm**.**
  - **ANN-Converted SNN** (High Precision, Large Delay)**:** avoids the difficulties faced by the directly trained SNN.

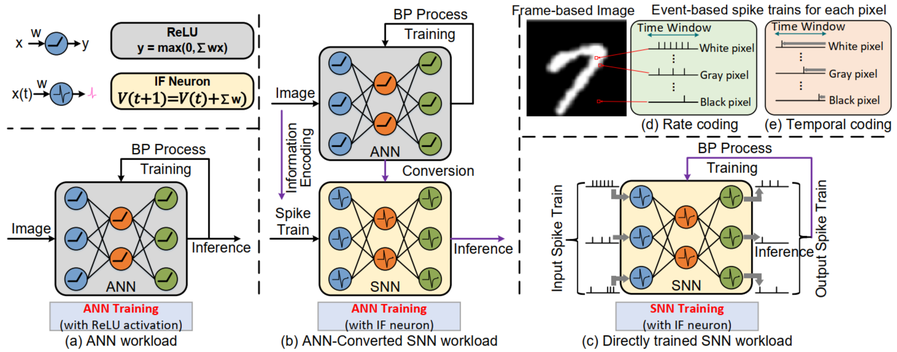


Illustration of the workload varying in SNN types

SNN COMPRESSION

- **Goal**: The compression of SNNs in this work targets the reduction of memory and computation in inference.

- There are generally two alternatives to reduce memory and computation:
  - **Connection pruning:** reduces the number of synapses between neurons.
  - **Weight quantization:** reduces the bit-width of the synapse.

- These methods learn from the existing compression techniques in the field of ANNs and directly apply them to SNNs.

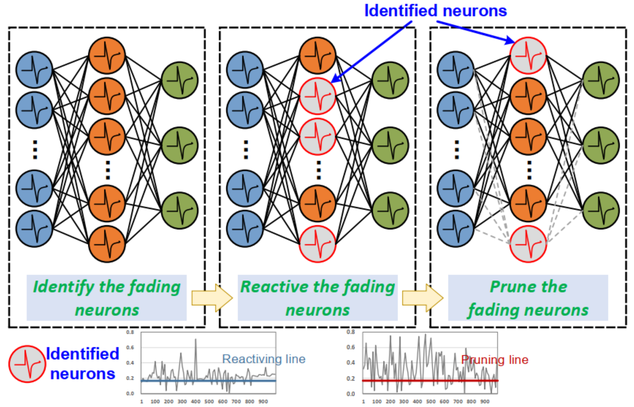
DYNSNN FRAMEWORK

Motivations

- There is an extra opportunity to reduce the compute cost according to the neuron activity.

Bio-inspired Pruning Framework

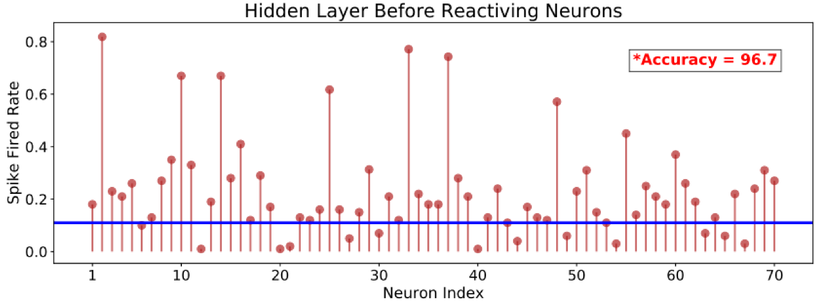
- We selectively mask some useless or even negatively acting neurons during the propagation of SNNs. The reduction in the number of neurons results in significant improvement in the efficiency of the SNN inference phase
- This is because neurons are the basic computational units of SNN, enabling decreasing the number of neurons directly reduces the need for PE in SNN implementation.
- We identify the fading neurons (i.e., with the low neuron activity) by the threshold determined during training and mask off these neurons from the SNN computation.



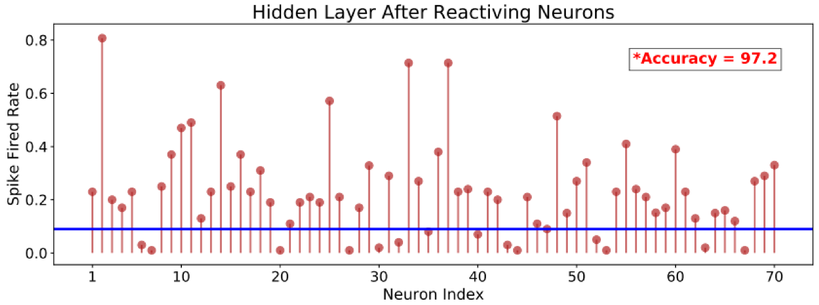
Schematic of DynSNN applied in the SNN training process

EVALUATIONS

The change in firing rate value before and after applying DynSNN to neurons in the hidden layer.



(a)



(b)

Performance Conparsions on MNIST

Pruning Methods	SNN Type	Arch.	Acc. (%)	Acc. Drop(%)	Comp. Rate
Deep R [36]	SNN training	LSNN	93.70	+2.70	88%
ADMM [37]	SNN training	LeNet-5 like	99.07	-0.43	60%
Grad R [35]	SNN training	3 FC	98.92	-0.33	74.29%
this work	**SNN training**	**3 FC**	**99.23**	**-0.02**	**57.4%**
this work	**SNN training**	**3 FC**	**98.98**	**-0.27**	**69.7%**
this work	**ANN-to-SNN**	**LeNet**	**99.15**	**-0.35**	**61.5%**