ATTACK ON SNN NETWORK AND DEFENSIVE MEASURES

Agasthya Harekal

201 Vairo Blvd, State college, PA

+1 8148269864

adh5677@psu.edu

Abstract

This paper summaries attack on Spiking Neural Network and defensive measures. The dataset considered is the MNIST dataset for classification. Here, we consider attacks such as firing threshold and spike timing manipulation/distortion based attacks. The paper also considers defensive techniques which are required to decrease the effect of attack on spiking neural network.

1. Introduction

Some Deep Neural Networks are historically brain-inspired, but there are substantial differences in the structure, neural computations, and learning rule compared to the brain. The spiking neural network are artificial neural networks which are third generation of neural network models that have emerged rapidly as a better design option when compared to the traditional deep neural network since inherent model structure and properties that is similar to the brain's functionality. Some of the advantages of SNNs include lesser number of neurons are required to realise same computations, higher energy efficiency because they process information when spike arrives and spike events are sparse in time. Also, spiking neurons have striking similarity to the biological ones because they use discrete spikes to compute and transmit information and hence they are highly sensitive to temporal characteristics of processed data. Deep SNNs provide appropriate architectures for developing an efficient, brain-like representation. Also, pattern recognition in the primate's brain is done through multi-layer neural circuits that communicate by spiking events.

An SNN architecture consists of spiking neurons and interconnecting synapses that are modeled by adjustable scalar weights. The first step in implementing an SNN is to encode the analog input data into the spike trains using either a rate based method or some form of temporal coding, or population coding. A biological neuron in the brain (and similarly in a simulated spiking neuron) receives synaptic inputs from other neurons in the neural network. In comparison to true biological networks, the network dynamics of artificial SNNs are highly simplified. In this context, it is useful to assume that the modeled spiking neurons have pure threshold dynamics. The activity of pre-synaptic neurons modulates the membrane potential of postsynaptic neurons, generating an action potential or spike when the membrane potential crosses a threshold. A model of action potential generation was created from voltage gaining properties of the ion channels in the squid cell membrane of the squid axon.

Spike trains in a network of spiking neurons are propagated through synaptic connections. A synapse can be either excitatory, which increases

the neuron's membrane potential upon receiving input, or inhibitory, which decreases the neuron's membrane potential. The strength of the adaptive synapses (weights) can be changed as a result of learning. The learning rule of an SNN is the most challenging component for developing multi-layer (deep) SNNs, because the non-differentiability of spike trains limits the popular back-propagation algorithm. Learning rules in SNNs is realised by adjusting scalar valued synaptic weights. Spiking enables a type of bio-plausible learning rule that cannot be directly replicated in non-spiking networks. The learning rule can be divided into unsupervised learning via STDP, probabilistic characterisation of unsupervised STDP.

Deep Learning is implemented in SNNs using STDP and stochastic gradient descent. Various types of spiking deep learning approaches are deep fully connected SNNs, spiking CNNs, spiking DBNs, and spiking RNNs.

Next, we think of attacks on Spiking Neural Networks. We understand there are various types of adversarial attacks that include FGSM attack, IFGSM attack, MI-FGSM attack, Deep-fool attack, one-pixel attack. We understand the differences between the various attacks. Other type of attack is efficiency based attack.

The dataset considered for research is the MNIST dataset where experiments are conducted such as attack on the SNNs and how the defensive measures are able to decrease the effectiveness of the attacks. The MNIST database contains 60000 training images and 10000 test images. The MNIST dataset is a large database of handwritten digits that is commonly used for training various image processing systems. It was created by remixing the samples from NIST's original datasets.

2. Attacks and Defensive Measures

As part of research we conduct attacks on the SNN neural network and get results. The dataset considered is the MNIST dataset for classification. The SNN network presents an accuracy of 74.83% percent for training data and 86% percent for test data. The SNN network classifies the images of MNIST dataset with good accuracy so that experiments can be conducted on effect of attacks on SNN network and how the defensive measures can decrease the extent of an attack.

First, we consider the attacks on the network. It is understood that attacker is a person who is going to attack the network by manipulating the input parameters of the network. Here, we consider that the attacker manipulates two important parameters of the SNN network that is firing threshold and the spiking timing. We consider positive and negative percentage change in both the parameters and accurately record the change in the accuracy of classification for training data and test data, thereby assessing the extent of attack. Here, the percentage change considered is in the steps of 0%, 5%, 10 %, 20%, 50%, 100% and so on.

As part of the defensive measures are concerned, after evaluation of various parameters, two parameters start_inhibition, max_inhibition were considered to be found effective in decreasing the extent of the attack. We accurately recorded the accuracy of the classification after these parameters were set and the network came under attack.

3. Results

As mentioned above elaborate evaluation of decrease in the accuracy due to attack was done. Here, we analyse the effect of positive and negative percentage change for the parameter firing threshold and it was found that for change in the parameter there was no change in the accuracy for training and test data until a 50% positive change was done and for negative change, there was no change in the accuracy apart from fluctuations for training data and test data. For positive change of 50%, the training, test data accuracy decreases to 64%, 82% and for 150 % the accuracy decreases to 25.37%, 49% for training and test data respectively.

We analyse change in accuracy for spike timing difference where the parameter was increased or decreased at the scale of 10. For positive change, we realise that there is no change in the accuracy and the accuracy was set at 60.37%, 79% for training and test data, although the progress towards final accuracy was different since accuracy at 10%, 15% , 20% and so on were different. For negative change in spike timing the accuracy decreases. At 10^3 change, the accuracy is at 22.91%

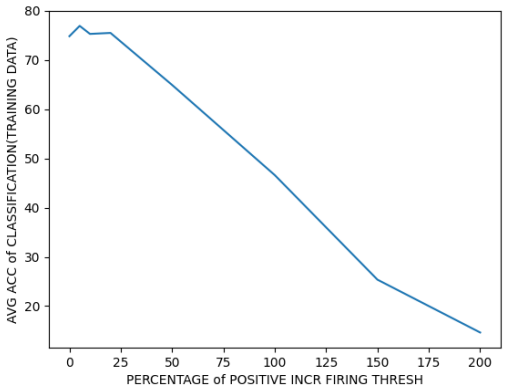

Fig 1: Change in avg accuracy due to positive increase in the firing threshold for training data.

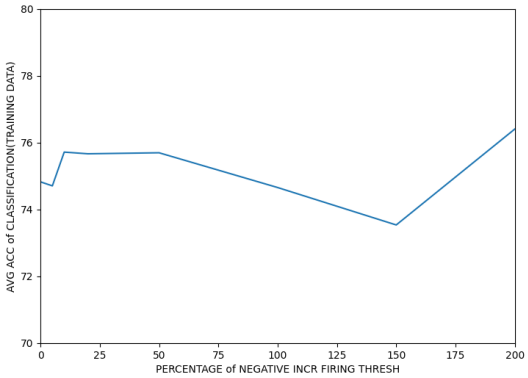

Fig 2: Change in avg accuracy due to negative increase in firing threshold for training data.

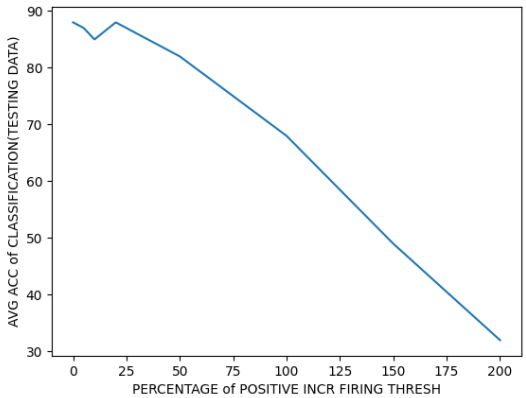

Fig 3: Change in avg accuracy due to positive increase firing threshold for test data.

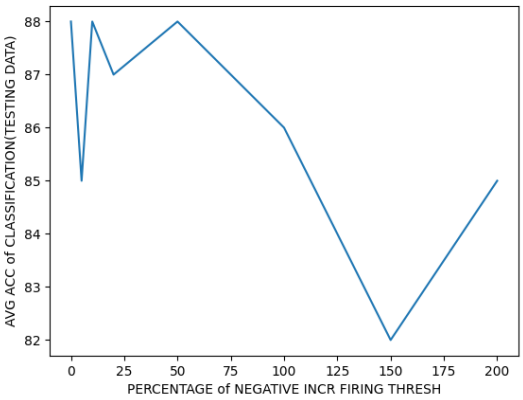

Fig 4: Change in avg accuracy due to negative increase of threshold for test data.

Further, we get results for defensive measures. For threshold=26.0, the accuracy was found to increase from 25.37%, 49% for training data, test data to 48.3%, 77% when max_inhibition=30. When spike_timing=1e3, start_inhibition=30, the accuracy increases from 14.67%, 16% to 28.91%, 37% and increases to 31.19%, 25% when max_inhibition=0.

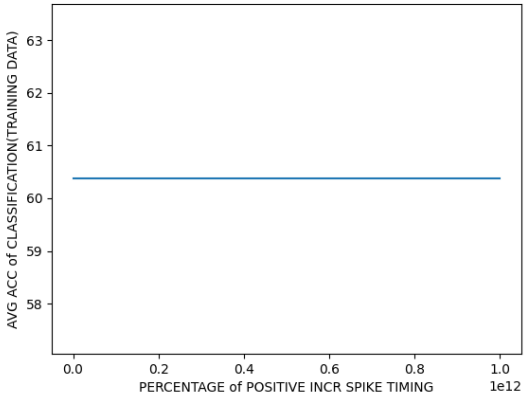

Fig 5: Change in avg accuracy due to positive increase of spike timing for training data.

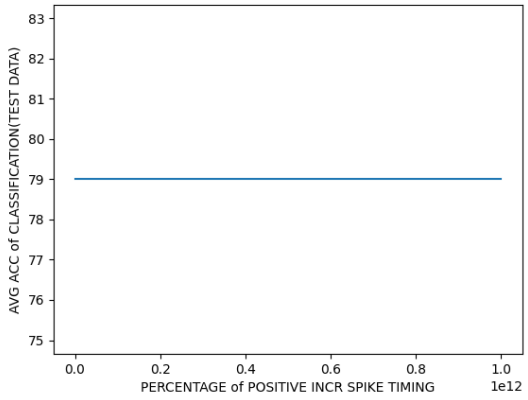

Fig 6: Change in avg accuracy due to positive increase of spike timing for test data

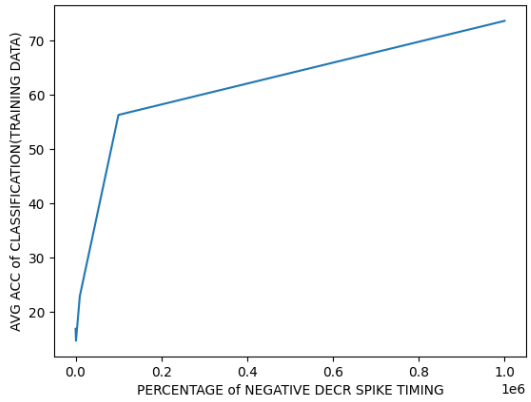

Fig 7: Change in avg accuracy due to negative increase of spike timing for training data.

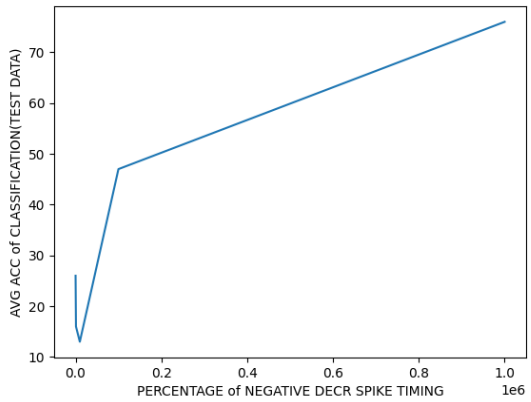

Fig 8: Change in avg accuracy due to negative increase of spike timing for test data.

	Avg Accuracy	
percentage change	Positive Difference	Negative Difference
0%	74.83%	74.83%
5%	76.93%	74.71%
10%	75.30%	75.72%
20%	75.50%	75.67%
50%	64.93%	75.67%
100%	46.61%	74.66%
150%	25.37%	73.54%
200%	14.62%	76.41%

Fig 9: Table above shows change in avg accuracy due to positive and negative increase of firing threshold for training data.

	Avg Accuracy	
percentage change	Positive Difference	Negative Difference
0%	86%	86%
5%	87%	85%
10%	85%	88%
20%	88%	87%
50%	82%	88%
100%	68%	86%
150%	49%	82%
200%	32%	85%

Fig 10: Table above shows change in avg accuracy due to positive and negative increase of firing threshold for test data.

	Avg Accuracy	
10^n change	Positive Difference	Negative Difference
10^1	60.37%	73.69%
10^2	60.37%	56.33%
10^3	60.37%	22.91%
10^4	60.37%	14.67%
10^5	60.37%	16.88%

Fig 11: Table above shows change in avg accuracy due to positive change and negative increase of spike timing for training data.

	Avg Accuracy	
10^n change	Positive Difference	Negative Difference
10^1	79.0%	76.0%
10^2	79.0%	47.0%

10^3	79.0%	13.0%
10^4	79.0%	16.0%
10^5	79.0%	26.0%

Fig 12: Table above shows change in avg accuracy due to positive change and negative increase of spike timing for test data.

The tables have the recorded data in detail for change in accuracy of training and test data for percentage change in parameters.Tables also depict improved accuracy for training and test data using defensive measures when the network is under attack.

	Accuracy at firing threshold=26.0	
max_inhibition	training data	Test data
-20	46.97%	71%
60	36.93%	67%

Fig 13: Table above shows improved accuracy for training and test data when max_inhibition is set at different values when firing threshold=26.0.

	Accuracy at firing threshold=26.0	
start_inhibition	training data	Test data
20	53.68%	81%
30	48.3%	77%

Fig 14: Table above shows improved accuracy for training and test data when start_inhibition is set at different values when firing threshold=26.0

	Accuracy at spike timing=1e3	
max_inhibition	training data	Test data
-20.0	40.35%	34%
0.0	31.19%	25%

Fig 15: Table shows improved accuracy for training and test data when max_inhibition is set at different values when spike timing=1e3.

	Accuracy at spike timing=1e3	
start_inhibition	training data	Test data
20	37.37%	32%
30	28.91%	37%

Fig 16: Table shows improved accuracy for training and test data when start_inhibition is set at different values when spike timing=1e3.

4. Conclusion

In this paper, we discussed about Spiking Neural Networks, its architecture, its implementation of deep learning based concepts and also the attacks on the SNN networks. We also discussed about MNIST dataset.
Discussion was done about attacks on SNN network that was trained for MNIST data classification. Here, graphs are used to depict the change in the accuracy as the parameters are modified. Also discussion was done on the defensive measures that can be taken to improve the accuracy.

References

1) Hananel Hazal, Daniel Saunders, Darpan T Sanghavi, Hava Siegelmann, Robert Kozma. 2018.“Unsupervised Learning with Self Organizing Spiking Neural Networks”. 2018 International Joint Conference on Neural Networks (IJCNN).

2) Albert Marchisio, Giorgio Nanfa, Faiq Khalid, Muhammad Abdullah Hanif, Maurizio Martina, Muhammad Shafique.2019.“SNN under Attack: are Spiking Deep Belief Networks vulnerable to Adversarial Examples?”. 2020 International Joint Conference on Neural Networks (IJCNN).

3) Hananel Hazan, Daniel J Saunders, Darpan T Sanghavi, Hava Siegelmann, Robert Kozma. 2019. “Lattice map spiking neural networks (LM-SNNs) for clustering and classifying image data”. Springer Nature Switzerland AG 2019.

4) Amirhossein Tavanaei, Masoud Ghodrati, Saed Reza Kheradpiseheh, Timothee Masquelier, Anthony Maida. 2019.“Deep Learning in Spiking Neural Networks”. arXiv:1804.08150.

5) Karthikeyan Nagarajan, Junde Li, Sina Sayyah, Sachhidh Kannan.2022.“Fault Injection Attacks in Spiking Neural Networks and Countermeasures”. 10.3389. Frontiers in Nanotechnology.

6) Jilles Vreeken.“Spiking neural networks, an introduction”. Adaptive Intelligence Laboratory, Intelligent Systems Group, Institute for Information and Computing Sciences, Utrecht University.

7) Kashu Yamazaki, Viet Khoa Vo Ho, Darshan Bulsara, Ngan Le.2022.“Spiking Neural Networks and Their Applications: A Review”. National Library of Medicine.