Constructing An Associative Memory System Using Spiking Neural Network

A Seminar Report

submitted by

SANKAR VINAYAK E P

PKD19CS046

to

the APJ Abdul Kalam Technological University

in partial fulfillment of requirements for the award of degree

of

Bachelor of Technology

in

Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GOVERNMENT ENGINEERING COLLEGE PALAKKAD

SREEKRISHNAPURAM 678 633

2022 - 23

DEPT. OF COMPUTER SCIENCE ENGINEERING GOVERNMENT ENGINEERING COLLEGE PALAKKAD

2022 - 23

CERTIFICATE

This is to certify that the report entitled **Constructing An Associative Memory System Using Spiking Neural Network** submitted by **SANKAR VINAYAK E P** (PKD19CS046), to the APJ Abdul Kalam Technological University in partial fulfilment of the B.Tech. degree in Computer Science and Engineering is a bonafide record of the seminar work carried out by him under my guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Liji L Dominic
(Seminar Guide)
Assistant Professor
Dept.of CSE
GOVERNMENT ENGINEERING
COLLEGE PALAKKAD

Dr. Swaraj K P
(Seminar Coordinator)
Associate Professor
Dept.of CSE
GOVERNMENT ENGINEERING
COLLEGE PALAKKAD

Dr. Sabitha S
Professor and Head
Dept.of CSE
GOVERNMENT ENGINEERING COLLEGE
PALAKKAD

DECLARATION

I SANKAR VINAYAK E P hereby declare that the seminar report **Constructing An Associative Memory System Using Spiking Neural Network**, submitted for partial fulfillment of the requirements for the award of degree of Bachelor of Technology of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me under supervision of Liji L Dominic

This submission represents my ideas in my own words and where ideas or words of others have been included, I have adequately and accurately cited and referenced the original sources.

I also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

Sreekrishnapuram

31-12-2022

SANKAR VINAYAK E P

Abstract

An artificial neural network that can remember associations between input and output patterns is known as an associative memory system. Spiking neural networks, on the other hand, are a subset of artificial neural networks that imitate the firing of individual neurons over discrete time steps to model the behaviour of brain neurons. Combining these two concepts can result in an effective memory representation technique in which the contents can be accessed with speed and efficiency. The report provides an overview of the principles of associative memory and spiking neural networks, and then describes the architecture and training procedure for the system. The results show that spiking neural networks can be effective for implementing associative memory systems, and have potential applications in a range of computational neuroscience and machine learning problems.

Contents

List of Figures

Abbreviation

Abbreviation	Description
NN	Neural network
CAM	Content Addressable Memory
ANN	Artificial neural network
RNN	Recurrent neural network
STDP	Spiking-time-dependent plasticity
LIF	Leaky integrate and fire
SNN	Spiking neural network

Chapter 1

Introduction

The ability to store and retrieve associations between different stimuli is a fundamental component of many cognitive processes, including perception, learning, and memory. Associative memory is a type of memory system that allows for the storage and retrieval of information based on the relationships between different items in memory. It is a key component of a good deal of artificial intelligence and machine learning systems and has been extensively studied in both neuroscience and computer science. According to research done by google for fast contextual adaptation of speech [1] associative memory system using ANN are efficient in case of contextual adaptation of information.

Spiking neural networks is a kind of neural network which is capable of simulating the dynamics of individual synapses and neurons in the human brain. They have potential applications in many different domains, including computational neuroscience and machine learning, etc. They are efficient for simulating a variety of cognitive and sensory processing tasks of biological neural networks.

This work presents a study on generating an associative memory system using a spiking neural network. This study discusses the architecture and training process of the system and evaluates its performance as an associative memory. It also examines the potential applications of SNNs in implementing associative memory systems and also in the fields of computational neuroscience and machine learning.

Chapter 2

Literature Review

2.1 Content addressable Memory

Content addressable memory also known as Associative memory , was first proposed in the early 1960s by researchers at IBM, including Richard W. Harker, Kenneth C. Thompson, and Robert D. Denny. Using the content of the data as the address, CAM is a form of computer memory system that enables quick searching and retrieval of data.

In traditional random access memory (RAM), data is stored in a specific location based on a numerical address and the data can be retrieved by accessing the corresponding address. In contrast, CAM stores data in a specific location based on the content of the data and the data can be retrieved by searching for the specific content rather than the numerical address.

CAM has several advantages over traditional RAM, including faster search and retrieval times and the ability to store and retrieve data based on the content rather than a numerical address. These features make CAM particularly useful in applications where rapid searching and retrieval of data is important, such as database management and pattern recognition.

The concept of CAM has had a significant impact on the field of computer science. The work of Harker, Thompson and Denny has contributed to the development of efficient and effective methods for storing and retrieving data in computers. Its applications include database management systems which require searching through the data as fast as possible figure2.1 shows one such circuit.

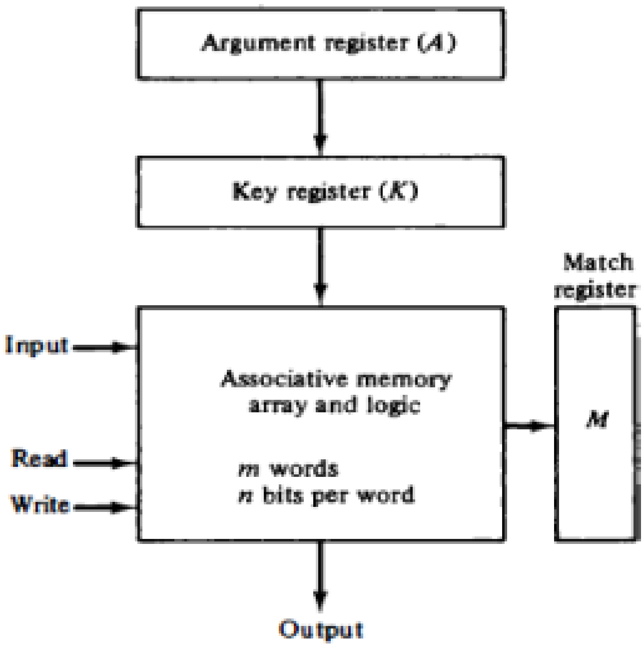

Figure 2.1: Associative memory circuit

2.2 Associative network

A particular kind of artificial neural network called an associative network is made to store and retrieve data based on the strength of the connections between neurons. Associative networks are frequently employed for tasks like image recognition, natural language processing, and recommendation systems that demand the recall of information based on the relationship and pattern.

2.2.1 Influence of Context on Letter Perception:An Interactive Activation Model

Introduction

The interactive activation model [2] is a computational model that attempts to explain how the brain processes and interprets written language. According to this theory, the brain stores and retrieves information about letters and words based on the strength of the connections made between neurons. The bottom-up and top-down processes have different impacts on the strength of these connections.

The model proposes that the brain has a hierarchical structure, with lower-level processing units representing features such as lines and curves, and higher-level processing units representing letters and words. According to the model, the activation of a processing unit depends on both the activation of its inputs and the activation of its outputs.

Proposed system

The authors propose a hierarchical structure for the model, with lower-level processing units representing features such as lines and curves, and higher-level processing units representing letters and words. They contend that a processing unit's activation is reliant on the activations of both its inputs and outputs. The system's two primary ideas are

Auto associative memory: A single layer NN with the same number of training and output vectors. The weights are determined by the patterns to be stored.

Hetero associative memory: A single-layer NN with different numbers of input training vectors and output vectors. The pattern stored in the network determines the weights. Because it is static, there will be no linear or delay operations.

Advantages

The model has the advantage of providing a computational framework for understanding how the brain processes and interprets written language. It suggests a hierarchical structure for the brain, with lower-level processing units representing features such as lines and curves, and higher-level processing units representing letters and words. Another advantage of the model is that it is based on the concept of auto-associative and heteroassociative memory, which are thought to be involved in how the human brain processes language and perception. As a result, the model is able to capture the intricate connections between letters and words and explain a variety of occurrences in written language processing.

Disadvantages

A disadvantage of the model is that it is a simplified model of the brain, and it may not capture all the complexity and nuance of language processing. It is also based on a set of assumptions and simplifications, and it may not accurately reflect the underlying mechanisms of the brain. In addition, the model is based on a set of equations and learning rules that are used to simulate the activation and deactivation of processing units, and these equations may not mirror the brain's internal workings precisely.

2.2.2 Neural networks and physical systems: The Hopfield network

Introduction

The Hopfield model [3] is a type of RNN that is designed to mimic the behaviour of neurons in the brain. It is a type of associative memory system and hence it can store and recall information based on the relationship between the data stored. It has applications including pattern recognition, optimization, and error correction.

Methodology

The use of computational simulations and models to analyse the behaviour of distributed systems, including neural networks, is covered in this article. He suggested the Hopfield network, a sort of associative memory network. A Hopfield network is made up of a collection of connected neurons arranged in a single layer. Each neuron in the network is fully interconnected, and the strength of the connections between neurons is dependent on the interactions between inputs and outputs. A collection of equations that describe the activation and deactivation of the neurons over time dictate the behaviour of a Hopfield network. These equations are based on the idea of auto-associative memory, where the original input can be retrieved from the output.

Advantages

The Hopfield model's relative simplicity and ease of implementation are two benefits. It is made up of a single layer of fully linked neurons, and the strength of the connections between them is determined by the interactions between inputs and outputs. This simplicity makes the Hopfield model easy to understand and implement. Another advantage of the Hopfield model is that it is depending on the strength of the connections between neurons, is that it is capable of storing and retrieving patterns or sequences of data. This makes it useful for tasks that require the recall of specific patterns or sequences of data, such as systems for recommendation, natural language processing, and picture identification. Also, this model is relatively robust and resistant to noise. These advantages make it useful in applications like image or speech recognition systems.

Disadvantages

One disadvantage of the Hopfield model is that it is relatively simple and may not be able to capture the complexity and nuance of more advanced neural networks. Another drawback is that it can retrieve the information only with the original input. A third disadvantage of the Hopfield model is that it can be sensitive to initialization and may not always converge to a stable state when used with a large amount or insufficiently distinct data. This can make it difficult to use the model for certain types of tasks, such as optimization or decision-making.

2.3 Spiking neural network

SNN are a subset of neural networks that mimic the behaviour of actual neurons by encoding and transmitting information via spikes or pulses. They are a relatively new variety of neural networks with the potential to raise the effectiveness and performance of artificial intelligence systems. Building associative memory systems using spiking neural networks is a relatively new field of study that has only lately begun to attract attention.

2.3.1 Training strategy for spiking neural networks for classification: SWAT

Introduction

SWAT [6](Spiking Weight association Training) is a method for training spiking neural networks for classification tasks. To enhance SNN performance, it combines supervised and unsupervised learning. Its foundation is the notion that the network's input and output patterns can be used to modify the weights of the connections between its neurons. The error between the network's actual output and the desired output is taken into consideration when adjusting the weights using a learning rule.

Methodology

They propose a method which merges Bienenstock-Cooper-Munro learning rule (BCM) [9] with Spike Time Dependent plasticity Spike-timing-dependent plasticity (STDP) [8].The result is a weight distribution that is unimodal, with weights and STDP being correlated.

The BCM learning rule is based on the notion that the network's connections between neurons should be strengthened in accordance with the activity of those neurons and the discrepancy between the network's actual output and the anticipated output. The rule states that if the activity of the neurons is high and the error is also high, the weights of the connections between neurons should be increased, and if the activity of the neurons is low and the error is also low, the weights should be decreased.

For neuromorphic computing, the research that is influenced by the structure and operation of the brain, STDP is an unsupervised learning strategy based on how neurons work in the brain. Based on the relative timing of spikes or impulses, the strength of the connection between neurons fluctuates during the process. The fundamental tenet of STDP is that if two N_{prede} and N_{succe} neurons are connected, and their spike time are w_1 and w_2 respectively according to STDP

Strength of connection between N_{prede} to N_{succe} should increase, if $\boldsymbol{w_1 w_2}$

Strength of connection between N_{prede} to N_{succe} should decrease, if $\boldsymbol{w_1 < w_2}$

Strength of connection between N_{prede} to N_{succe} should remain same, if $\boldsymbol{w_1 = w_2}$

In general, the foundation of SWAT is the notion that the network's input and output patterns can be used to modify the strength of connections between neurons in the network. The error between the network's actual output and the desired output is taken into consideration when adjusting the weights using this rule.

Advantages

One of the main advantages of SWAT is that it spiking neural networks for classification problems can be trained using this technique. Since SNN are a promising sort of artificial neural network that is believed to have the potential to enhance performance, therefore understanding this method is crucial for better artificial intelligence systems. The network is trained using a mix of supervised learning and unsupervised learning. As a result, the network can both learn from tagged samples and independently find patterns and features in the data.It is reasonably simple to execute and comprehend since it is based on the straightforward learning rules of altering the weights of the connections between neurons in the network based on the input and output patterns of the network. Overall, it could make SNN use possible in a variety of applications.

Disadvantages

Sometimes this may become computationally intensive, especially for large networks with many neurons and connections. This could make it difficult to use SWAT on systems with limited computational resources. Another potential disadvantage is that it is based on supervised and unsupervised learning, when efficient network training for supervised learning necessitates a significant volume of labelled data. In some circumstances, getting this could be challenging. It also depends on the network's initial weights for connections between neurons, and the network's performance may be significantly impacted by these weights. In order to obtain good performance, SWAT may be sensitive to initialization and may need the initial weights to be carefully tuned.

2.3.2 Backpropagation for spiking neural network:SpikeProp

Introduction

Spiking neurons may be more biologically realistic and efficient than other types of artificial neurons,additionally, they might be better suited for specific jobs like pattern recognition and natural language processing. However, they also note that training networks of spiking neurons can be difficult due to the non-differentiable nature of the spiking function. To address this problem, the authors propose the use of SpikeProp [4], which is a method for training networks of spiking neurons using backpropagation.

Methodology

It is a technique for employing backpropagation to train networks of spiking neurons. The spiking function, which describes the connection between a spiking neuron's input and output, cannot be differentiated. This means that networks of spiking neurons cannot be trained using conventional techniques for artificial neural networks, such as backpropagation. But it can be used to train networks of spiking neurons while avoiding this problem by using a surrogate gradient function insted of the ordinary gradient ued in ANN. It approximates the gradient of the spiking function.

Advantages

It enables the training of networks of spiking neurons using backpropagation. This makes it possible to use backpropagation's ease of use and effectiveness for training spiking neural networks. Spiking neural networks can be built resembling more the biological systems than the conventional ANN and make it better suited for particular tasks. It is also more efficient in training SNN than other techniques. It makes it effective in applications like image recognition and language translation.

Disadvantages

It is not effective in all applications of spiking neural networks. Also, due to the approximations used in surrogate gradient function, it may not be always accurate which may affect its performance.

Chapter 3

Methodology

Construction of associative memory using SNN [10] in this method consist of four phases

1. Initialization: Initialization of SNN and the input spiking signals
2. Structure formation: New connections with neighbouring neurons are formed
3. Parameter training: Optimize weight of synapse based on STDP
4. Pruning: Removing unnecessary connections to improve efficiency

3.1 Initialization

It involves two sub-process in which the data is preprocessed and converted into spiking signals and the network is initialized.

3.1.1 Initialization input spiking signals

The input to the SNN is spiking signals for that the input values need to be converted into spiking signals. For the example purpose here MNIST dataset is used which contains handwritten characters on digits. The figure 3.1 shows the steps involved.

Four convolutional kernels of size 4×4 shown in figure 3.2 is applied to the picture pixel values to extract the features. The input image into the kernel is of size 28×28 and the convolutional kernels reduce the shape into 24×24. Next, these values are passed through a maximum pooling layer shape 2×2. It makes the image smaller dimension of 12×12. The conversion layer converts these values into spiking encoding. Pixel

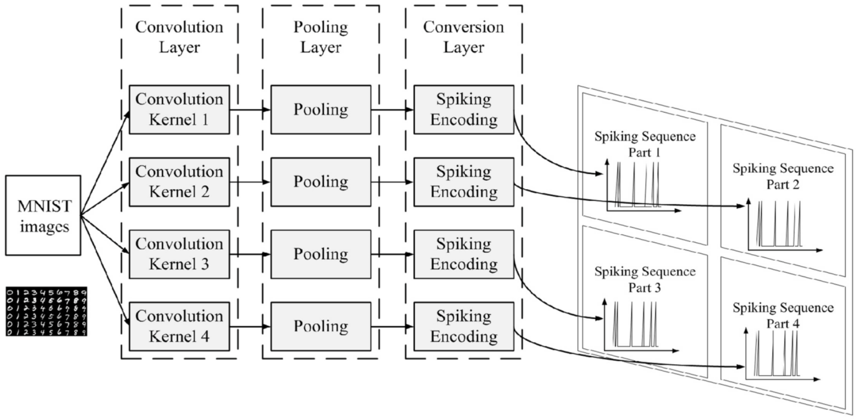

Figure 3.1: Data preprocessing

value in the range of [0,255] converted to delay in a spike from [0,100]ms. The delay is shorter if value higher. To cover the values first min-max normalization is used in

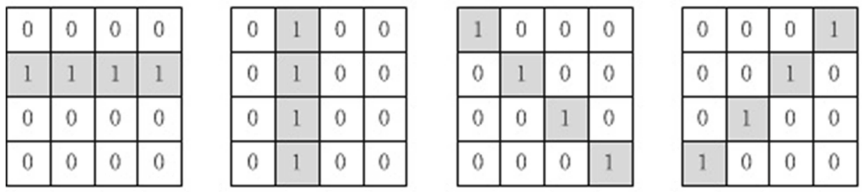

Figure 3.2: Kernels used

which, if the value of a pixel is d then,

$$N(d) = \frac{d - d_{\min}}{d_{\max} - d_{\min}}$$

Where $d_{\min}$ and $d_{\max}$ are maximum and minimum pixel values. Power encoding is used to get the spike time of pixel d,

$$S(d) = (T_{\max} - T_{\min}) \times (N(d) - 1)^2 + T_{\min}$$

where $T_{\min}$ and $T_{\max}$ are the spike's beginning and ending times.

3.1.2 Initialization of Spiking neural network

As shown in figure 3.3, the memory NN in this approach has three layers: input, memory, and output. The input layer receives the input of the spiking signal. For the neuron, the LIF model is used. The memory layer grows new connections to remember them. The process of producing the output is handled by the output layer. Both the number of input spiking signals and the number of neurons in the input and memory layers are equal which is 576. Ten neurons, which is equal to the number of output classes, make up the output layer.

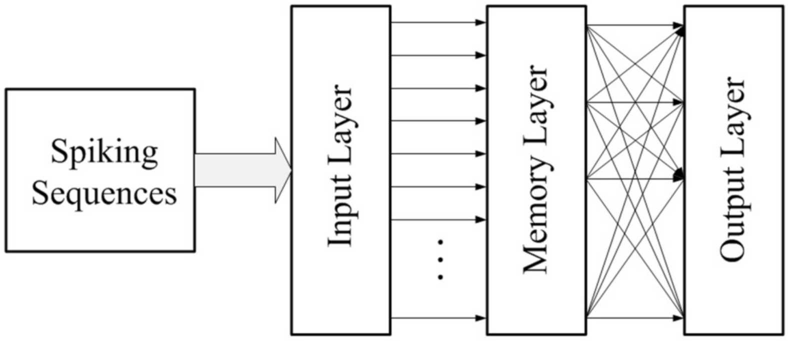

Figure 3.3: Structure of network

One-to-one connections are made between the input layer and the memory layer.

As a starting point, the synapse's weight is set at 50. It aids in eliciting the necessary reaction for Hebb's learning rule [7] to take place in the memory layer.

A coordinate value is given to each neuron in a layer. Using the spatial to temporal method depicted in the figure3.4, which encodes the spatial information of pixel values into the connection delay from the input layer to the memory layer. The delay in a neuron $i(x, y)$ connection where the pixel values in an image's x and y coordinates $p \times q$ input layer to the the corresponding neuron in the memory layer is calculated as

$$d_{im(x,y)} = 1 + p * x + y$$

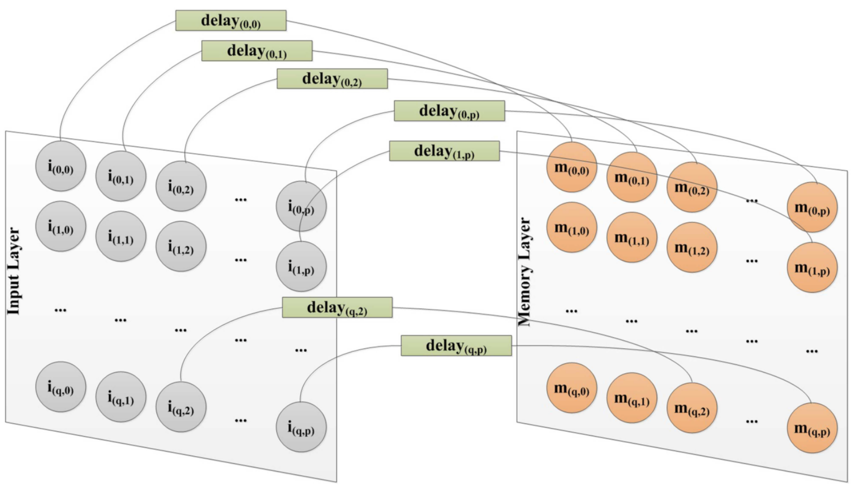

Figure 3.4: Delay in neuron connections

3.2 Forming network structures

Spiking signals are fed into the network during this phase. The behaviour of neurons in the memory layer is recorded, and new connections are made within the memory layer based on Hebb's learning rule [7]. New connection conditions are based on their threshold values of distance and time difference in firing two neurons to avoid explosive growth of connections. A new connection is made between two neurons if they meet the threshold requirements for firing delay and firing distance but do not already have one.The smaller the threshold lesser the number of connections that will be created.

Up until the stopping condition is met, it is repeated. Due to the LIF model's leaky properties, the connection between the memory layer and output layer uses the previously discussed spatial-to-temporal technique. The calculation of the connection delay between the memory layer and the output layer is

$$d_{mo(x,y)} = 1 + [N_m - delay_{im(x,y)}]$$

3.3 Parameter training

This phase is based on the ideas of STDP and reinforcement learning. It checks the recall ability of the network for the set of inputs. This phase does not change the weights of the connection between layers but rather changes the weight of the connection within the memory layer itself.The most often fired neuron in the output layer is taken into consideration as output for this procedure when a spiking sequence is supplied into the network. If the network could correctly recall, then there would be no need for optimization; otherwise, the weights would need to be changed. It works based on the following algorithm

Step 1: Pick an input image

Step 2: Feed input to the network

Step 3: Go to step 1 if the output layer's result is accurate; otherwise, go to step 4.

Step 4: Determine which neurons in the output layer S_O and memory layer S_M are firing inappropriately.

Step 5: If i is a neuron in the set S_M and j is a neuron in the set S_O and the strength of the connection between them is $W_{i,j}$, then $W_{i,j} = Shrink\_Coeff * W_{i,j}$

This process repeated for all the images. The value of $Shrink\_Coeff$ is constant between 0 and 1. Figure 3.5 demonstrates the memory layer's firing behaviour after receiving an input image corresponding to the number six. Colours indicate the time at which the spike occurred and lines indicate the connection between neurons.

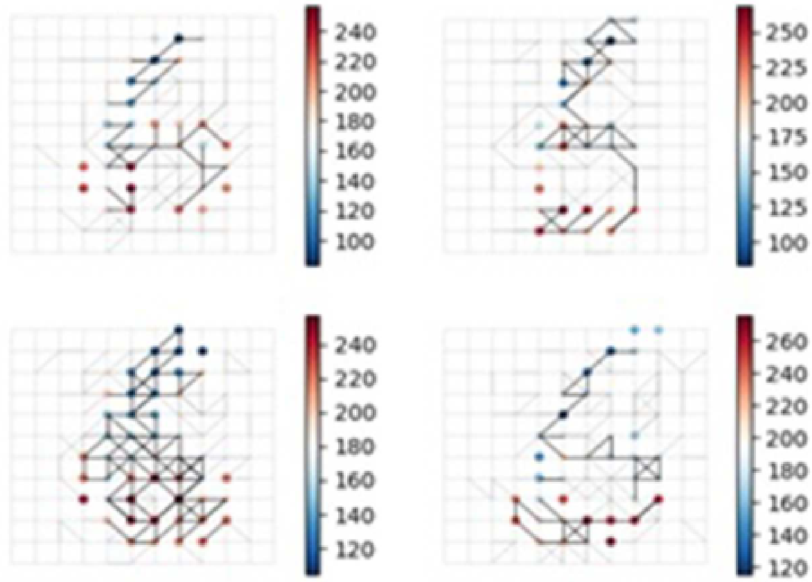

Figure 3.5: Recall response for number 6

3.4 Pruning

This phase helps in improving the efficiency of the network. If a connection's weight falls below the threshold during this period (here 3), the connection will be removed. The connection from the input layer to any memory layer neurons that don't have any connections to the output layer will likewise be eliminated.

Chapter 4

Results and Discussion

4.1 Development of the memory layer

The memory layer of the network did not have any connections between its neurons during the initialization phase, but new connections began to develop during the period of structure construction. To memorise various patterns, this network could develop various connections in the memory layer. The memory could recall more information with more free space available when the threshold was set to smaller values since the connection was more sparse.

4.2 Recall process

All of the photographs are utilised to test the recall process as the building is being built. The identical photographs were then utilised during the parameter training phase and the trimming phase. There are different reactions in the memory layer depending on the four different kernels that are used and illustrated by four images. When a picture is sent into the network, different components of the picture trigger different firing actions in the network. As was already mentioned, the firing sequence at the output layer uses a majority voting technique to decide the outcome of a picture sent into the network. The outcomes demonstrate that the network was able to recollect the images it had learned. The network also display some association ability between the newly learned information and the information that had previously been memorised when it was given with an unseen but comparable image of digit.

Chapter 5

Conclusion

This paper provided a technique for building an SNN-based associative memory system. Associative memory and SNNs are used in this method to produce a neural network design that can store and recall information based on the associations between various stimuli.

There are active research going on to develop neuromorphic hardware which improves the efficiency of different operation involved in the processing of information using an SNN. The hardware Lohi [11] and software for that Lava [12] are developed by intel for this application. Software like **N**EST [13] and SpikeTorch [14] is open source tool for the SNN. As more and more research is done in this field, it could be the next-generation machine learning algorithm.

References

[1] Munkhdalai, Tsendsuren, et al. *"Fast Contextual Adaptation with Neural Associative Memory for On-Device Personalized Speech Recognition."* arXiv, 2021, https://doi.org/10.48550/arXiv.2110.02220.

[2] McClelland, J. L., & Rumelhart, D. E. (1981). *An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. Psychological Review*, 88(3), 375-407.

[3] Hopfield J. J. (1982). *Neural networks and physical systems with emergent collective computational abilities.* Proceedings of the National Academy of Sciences of the United States of America, 79(8), 2554–2558. https://doi.org/10.1073/pnas.79.8.2554

[4] Bohté, Sander M. et al. *"SpikeProp: backpropagation for networks of spiking neurons."* The European Symposium on Artificial Neural Networks (2000).

[5] G. Indiveri, *"A low-power adaptive integrate-and-fire neuron circuit,"* Proceedings of the 2003 International Symposium on Circuits and Systems, 2003. ISCAS '03., 2003, pp. IV-IV, doi: 10.1109/ISCAS.2003.1206342.

[6] Wade, J. J., McDaid, L. J., Santos, J. A., & Sayers, H. M. (2010). *SWAT: a spiking neural network training algorithm for classification problems.* IEEE transactions on neural networks, 21(11), 1817–1830. https://doi.org/10.1109/TNN.2010.2074212

[7] Morris R. G. (1999). D.O. *Hebb: The Organization of Behavior,* Wiley: New York; 1949. Brain research bulletin, 50(5-6), 437. https://doi.org/10.1016/s0361-9230(99)00182-3

[8] Shouval, H. Z., Wang, S. S., & Wittenberg, G. M. (2010). *Spike timing dependent plasticity: a consequence of more fundamental learning rules.* Frontiers in computational neuroscience, 4, 19. https://doi.org/10.3389/fncom.2010.00019

[9] Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). *Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex.* The Journal of neuroscience : the official journal of the Society for Neuroscience, 2(1), 32–48. https://doi.org/10.1523/JNEUROSCI.02-01-00032.1982

[10] He, H., Shang, Y., Yang, X., Di, Y., Lin, J., Zhu, Y., Zheng, W., Zhao, J., Ji, M., Dong, L., Deng, N., Lei, Y., & Chai, Z. (2019). Constructing an Associative Memory System Using Spiking Neural Network. Frontiers in neuroscience, 13, 650. https://doi.org/10.3389/fnins.2019.00650

[11] Intel Corporation. (n.d.). LOHI: Low Overhead High Throughput Inference. Retrieved from `https://www.intel.com/content/www/us/en/newsroom/news/intel-unveils-neuromorphic-loihi-2-lava-software.html`

[12] Intel Corporation. (n.d.). Lava: Low Overhead Accelerator for Vision Algorithms. Retrieved from `https://www.intel.com/content/www/us/en/artificial-intelligence/solutions/inference/lava.html`

[13] Eppler, J. M., Helias, M., Muller, E., Diesmann, M., & Gewaltig, M.-O. (2008). NEST (NEural Simulation Tool). Scholarpedia, 3(4), 1430.

[14] Cao, Y., & Paine, T. (2020). SpikeTorch: A PyTorch Library for Spiking Neural Networks. arXiv preprint arXiv:2012.13350.

Appendices

Hebb's learning rule: [7] When an axon of a cell A is sufficiently close to excite a cell B, and repeatedly and or persistently takes part in firing it some growth related process or metabolic changes takes place in one or both of the cell such that A's efficiency, as one of the cells firing B is increased.

Leaky integrate and fire [5]

$$V(t) = \begin{cases} \beta.V(t-1) + V_{in}(t) \;\; when \;\; V < V_{th} \\ V_{reset} and set spike \;\; when \;\; V \geq V_{th} \end{cases}$$