Improving the Multimodal Classification Performance of Spiking Neural Networks Through Hyper-Parameter Optimization

Jin Seon Park
Department of Computer Science and Engineering
Kyung Hee University, 446-701
Republic of Korea
jinseon72@khu.ac.kr

Choong Seon Hong
Department of Computer Science and Engineering
Kyung Hee University, 446-701
Republic of Korea
cshong@khu.ac.kr

***Abstract*—Spiking Neural Networks (SNNs) are computational models that emulate the spike-based communication found in biological neural networks. These models are increasingly recognized for their potential to process sensor data in a biologically analogous manner, particularly within multimodal contexts involving both image and audio data. Nonetheless, optimizing the classification performance of deep SNNs is a complex task, frequently impeded by the intricate interactions of hyper-parameters. This paper addresses this challenge by employing advanced hyper-parameter optimization techniques to enhance the classification efficacy of a multimodal SNN. Our work not only refines the performance of SNNs on heterogeneous data types but also elucidates the intricate dynamics between hyper-parameter configurations and classification accuracy within these networks.**

***Index Terms*—multimodal classification, spike, SNNs, hyper-parameter optimization**

I. Introduction

Spiking Neural Networks (SNNs) are a class of artificial neural networks inspired by the functionalities of biological neural systems [1]. Designed to replicate the dynamics of neuronal activity, SNNs uniquely encode and convey information via discrete events known as spikes, which propagates through an intricately connected web of neurons and synapses. With their capacity for low-power computation and parallel processing, SNNs have attracted considerable attention in recent years, particularly for their proficiency in handling data from a variety of sensors. These networks excel in energy efficiency and parallel processing capabilities, akin to the human brain's remarkable ability to integrate multisensory information—from visual to auditory—enabling it to tackle complex tasks [2].

However, achieving optimal classification performance in SNNs is a challenging task, largely due to the complex interplay of hyper-parameters that govern the behavior of these models. Consequently, our research aims to identify the most effective combinations of hyper-parameters that can enhance the classification capabilities of SNNs, particularly when analyzing multimodal sensor data, including images and audio. We explore a suite of optimization techniques, including grid search, random search, and Bayesian optimization [3], among others, to determine their efficacy in tuning SNNs for superior performance.

To effectively manage the inherent complexity of multimodal data, our approach involves encoding each data modality into spike form for integration within the SNN architecture. The primary goal of this research is to closely emulate biological neural systems in SNNs, thereby devising an efficient technique for multimodal data processing. We anticipate that this biologically inspired approach will not only improve the accuracy of SNN-based classification models on multimodal datasets but will also broaden their practicality across various application domains.

II. Related Work

A. Spiking Neural Network

Spiking neural networks (SNNs) are designed to emulate the operational principles of biological brains and have been recognized for their ability to effectively process data from various sensors. These networks are particularly noted for their energy efficiency and parallel processing capabilities [1]. In SNNs, spikes serve as the unit of information, which propagate through a complex network of neurons and synapses. Unlike conventional deep learning networks that communicate through continuous tensors or floating-point values, SNNs operate on the principle of discrete events, signifying whether a spike has occurred at a particular moment in time within a specific neuron.

The most crucial component in the structure of SNNs is the neuron model, which defines the role of neurons and

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-01287-005, Evolvable Deep Learning Model Generation Platform for Edge Computing), in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00207816), and in party by Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea Government (MSIT) (Artificial Intelligence Innovation Hub) under Grant 2021-0-02068-003,*Dr. CS Hong is the corresponding author.

979-8-3503-3094-6/24/$31.00 ©2024 IEEE

how they function [4]. The occurrence of spikes is determined by differential equations representing various biological processes. Several neuron models capture various aspects of how neurons behave, and among them, we have chosen the leaky integrate-and-fire (LIF) model [5]. This model is used to describe the relationship between the current and voltage inside the neuron, explaining how the membrane potential changes through chemical and electrical processes when the neuron receives input current. The differential equation for the neuron model LIF is as follows:

$$\tau_\nu \frac{dV}{dt} = (E_r - V) + g_e(E_e - V) + g_i(E_i - V). \quad (1)$$

τ_ν represents the time constant, indicating the time it takes to update the membrane potential of the LIF neuron. A larger value of v means that the membrane potential of neurons changes more slowly, while a smaller value leads to faster changes. E_r represents the resting membrane potential of the neuron. E_e and E_i are the equilibrium potentials of the excitatory and inhibitory synapses, respectively. They indicate at what potential the neuron's membrane receives input when synapses are activated. g_e and g_i are the conductance of the excitatory and inhibitory synapses, respectively. Each of them arises through excitatory synaptic signals and inhibitory synaptic signals. V represents the current membrane potential of the neuron.

The leaky integrate-and-fire (LIF) neuron model processes information through a dynamic mechanism wherein presynaptic neurons, upon activation, transmit signals to postsynaptic neurons. At each discrete time step, the membrane potential of the neuron is updated to reflect incoming spikes, with each spike incrementally raising the potential. Once this potential ascends to a certain threshold voltage V_{th}, the neuron emits an output spike. Immediately after firing, the membrane potential is reset to a lower voltage V_{reset}, and the neuron enters a refractory period during which it is temporarily incapable of firing again, regardless of incoming spikes. This cycle allows the neuron to emulate the timing-based information processing seen in biological systems. When multiple neurons form a network, this spiking mechanism enables complex temporal patterns of activity to emerge. The intricate details of this process are illustrated in Figure 1, which depicts the step-by-step changes in membrane potential leading up to, and following spike generation.

III. System Model

A. Framework

This paper emphasizes the classification of multi-modal data using SNNs. Our framework proposes a method that involves transforming image and audio data into spike representations and processing them simultaneously through the SNNs model, thereby enabling effective classification of multi-modal data.

Figure 2 visualizes our proposed method. In our approach, image and audio data are transformed into spike representations in distinct ways. For image data, spikes are generated based on the pixels of the image, while for audio data, spikes are generated using methods that capture audio-related features such as Mel Frequency Cepstral Coefficients (MFCC). These spike-transformed data are input into the input layer of the SNNs for training.

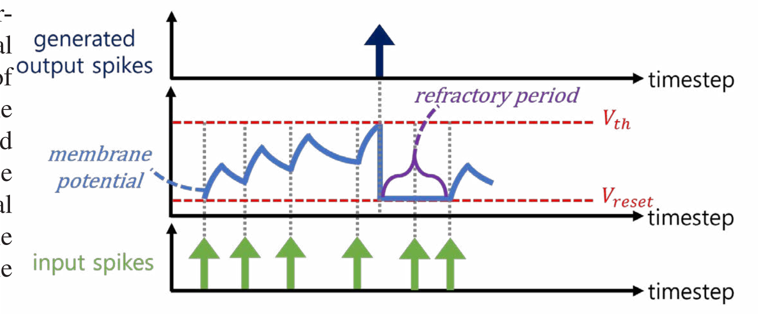

Fig. 1. Information processing mechanism of the LIF neuron model

The SNNs have been designed to effectively learn two different modalities of data. The image modality is input to the SNNs model through an input layer composed of a convolutional layer and a fully-connected layer, while the audio modality is input through a fully-connected layer that takes MFCC features as input. By using different input layers, we simultaneously learn and integrate the features of both modalities of data. We conducted research to optimize the hyper-parameters required by the SNNs to achieve the best classification performance.

B. Spike Encoding

To convert multimodal data into spikes, a data preprocessing process is required that converts each data's features into extractable forms.

First of all, for image data, changes such as increasing or decreasing brightness can be characterized by the image based on the pixels in the image and can be applied to the spike neuron model to be transformed into a spike. In this paper, the characteristics of the image were extracted by the difference between the bright and dark parts of the mnist image.

In the case of audio data, MFCC can be used to extract audio features and convert them into spikes[6]. MFCC is a numerical representation used to effectively represent and analyze audio signals, capturing the unique characteristics of sound. The process of extracting MFCC involves several steps. First, the audio signal is divided into frames, and the Fast Fourier Transform (FFT) is applied to each frame to obtain a spectrum. The audio signal is initially represented in the time domain, with time on the horizontal axis and sound pressure on the vertical axis. Applying FFT, an algorithm that transforms the signal into its frequency components, results in a representation in the frequency domain, known as the frequency spectrum. Mel filter bank is the process of applying a filter to obtain mel values for frequencies obtained for each frame. If you perform the Discrete Cosine Transform (DCT)[7] operation, which compresses and expresses the matrix for mel spectrum obtained earlier, the MFCC will come out as output.

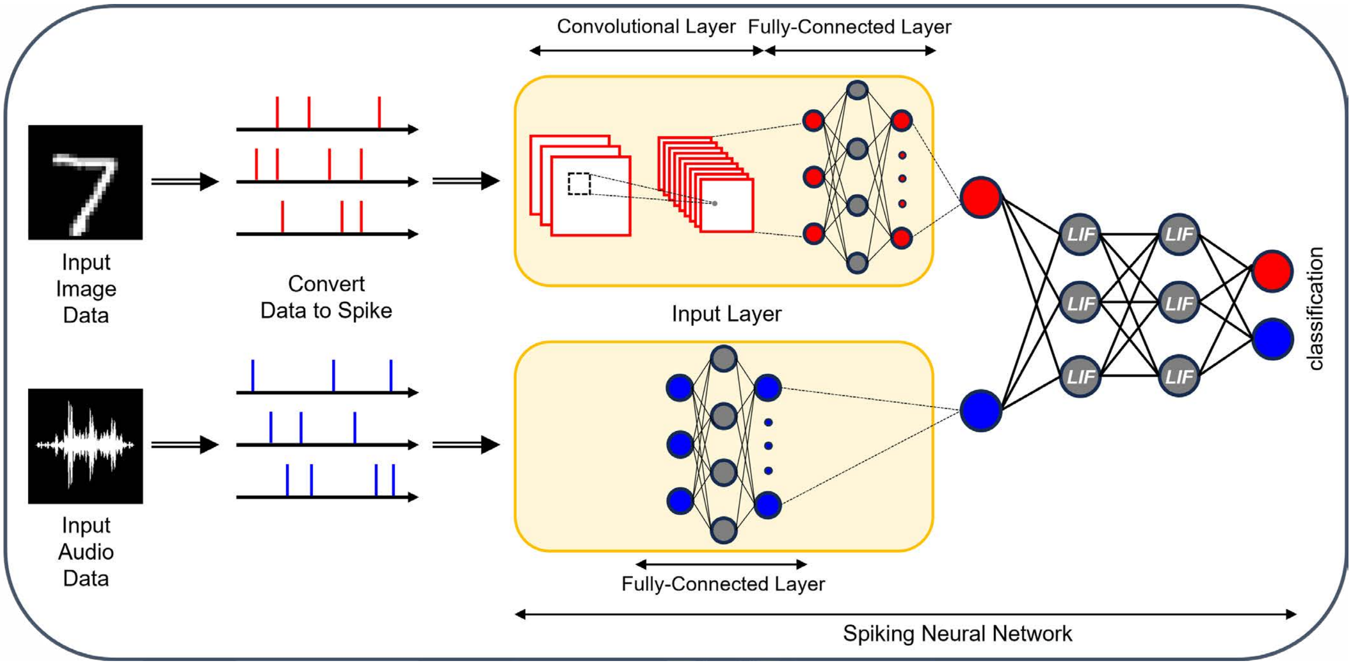

Fig. 2. Framework

The mel spectrum obtained in the previous process correlates with frequencies, and it plays a role in de-correlating this correlation using the log function in the DCT operation. The mathematical calculation of MFCC is represented in Equation (2), where m represents the frame number, n represents the sample index of the converted signal used in DCT. n is used to construct the MFCC coefficients obtained as a result of the DCT transformation. R represents the number of mel filters, the parameter of MFCC, and $MF_m[r]$ represents the mel filter bank value.

$$mfcc_m[n] = \frac{1}{R}\sum_{r=1}^{R} log(MF_m[r])cos[\frac{2\pi}{R}(r+\frac{1}{2})n]. \quad (2)$$

C. Mean Squared Error Spike Count Loss

We reconstruct the multimodal classification problem as a regression problem and use the mean squared error spike count loss function to do this. In general, the cross-entropy loss function is used a lot in the classification problem, which attempts to activate the correct class at every stage of time and prevent the wrong class from being activated at all. However, the Mean Squared Error (MSE) loss function learns to activate the correct class a given number of times over a given period, and allows the wrong class to be activated less than a given number of times [8]. In other words, the MSE loss function allows you to deal with classification problems by converting them into regression problems. The form of the MSE loss function is as follows:

$$MSE = \frac{1}{n}\sum(y-\hat{y})^2. \quad (3)$$

In Equation (3), y is the actual target value, $\hat{y}$ is the value predicted by the model. However, in this paper, we want to apply MSE to spike count, and we independently calculate and sum up the MSE loss for each output neuron by treating it as the target value for how much each neuron should generate spikes during a specific period. In this way, in our paper, learning is done by adjusting the rate of occurrence of spikes between the right and wrong classes.

IV. Experiment

A. Data

For the experiment, we used mnist image data and mnist audio data. For image data, we used mnist image dataset with labels ranging from 0 to 9. For audio data, we used mnist audio dataset consisting of voice numbers with labels from 0 to 9.

B. Hyper-parameter Optimization

Hyper-parameters are one of the key factors that greatly influence the learning and performance of deep learning models, especially for complex models such as SNNs, where the correct hyper-parameter setting plays an important role in determining classification performance. This paper introduces hyper-parameters and optimization techniques considered to improve classification performance.

One of the hyper-parameters considered in this paper to improve the classification performance of SNNs is the incidence rate of spikes. Consideration of mean squared error spike count loss is to set the rate of spike occurrence between the correct and incorrect classes as hyper-parameters. By adjusting the incidence rate of spikes, you can promote the occurrence of spikes for the correct class and suppress the occurrence of spikes for the wrong class. Spike incidence is adjusted during learning, and finding the optimal rate of spike is critical to improving classification performance. Another hyper-parameter considered is the learning rate that controls the rate of weight update. Too large learning rates hinder convergence, and too small learning rates can slow learning. Therefore, choosing the right learning rate plays an important role in SNNs learning process.

For hyper-parameter optimization, this paper utilized grid search, random search, and Bayesian optimization techniques.

Grid search is a way to find the best combination by dividing hyper-parameter space into several sections and trying all possible hyper-parameter combinations in each section. Grid search narrows the range of initial hyper-parameters and brings you closer to the optimal combination.

Random search[9] is a navigation method that randomly inputs hyper-parameter values and generates models using hyper-parameters that show superior values. It has the advantage of reducing computational costs while achieving good results.

Bayesian optimization [10] is a probabilistic optimization process that reduces unnecessary hyper-parameter iterations to quickly find the optimal hyper-parameter. It is a technique to find the optimal solution with the unknown objective function to the maximum (or minimum) and consists of an acquisition function and a surrogate model. The role of the acquisition function is to find the next most appropriate hyper-parameter candidate. Mathematically determine the next search point based on a model with probabilistic estimation for the objective function. Alternative models are probabilistic representations and models of objective functions. Bayesian optimization attempts optimization with probabilistic estimation.

By applying these optimization techniques, we focused on adjusting spike incidence and learning rates and improving the classification performance of SNN models. By finding the best hyper-parameter combination, we aim to improve the performance of SNNs in multimodal data classification tasks and increase their availability in a variety of applications.

V. Result

This section presents the results of classification experiments for image and audio data as unimodal sets, as well as the results for multimodal classification, and confirms the classification accuracy based on hyper-parameter settings. In multimodal classification, the best performance was achieved with a combination of hyper-parameters: a learning rate of 0.0001, a correction rate of 0.8, and a decay rate of 0.2. Random searches yielded the best results in each classification experiment, particularly achieving the highest accuracy in image classification. Performance differences were also observed between unimodal and multimodal classifications for image and audio data, respectively. Image classification demonstrated higher accuracy than audio classification, and multimodal classification achieved higher accuracy than audio unimodal classification. This suggests that multimodal approaches can improve classification performance by leveraging various types of input data. Due to SNNs' parallel processing capabilities, they can effectively integrate information from both modalities. In all cases, random searches achieved the highest accuracy, indicating that this method can effectively navigate the complex properties of multimodal data. Random searches discover the best combination by randomly testing different hyper-parameter combinations, which often leads to improved performance. Tables 1 and 2 below show the optimization results for unimodal and multimodal data, respectively.

TABLE I
Unimodal optimization RESULT

	Image modal		Audio modal	
method	hyper-parameter	accuracy	hyper-parameter	accuracy
Grid Search	Learning rate = 0.001 Correct rate = 0.7 Incorrect rate = 0.3	92.97	Learning rate = 0.02 Correct rate = 0.7 Incorrect rate = 0.3	81.22
Random Search	Learning rate = 0.01 Correct rate = 0.8 Incorrect rate = 0.2	**94.16**	Learning rate = 0.0001 Correct rate = 0.8 incorrect rate = 0.2	**86.64**
Bayesian Optimization	Learning rate = 0.02 Correct rate = 0.8 Incorrect rate = 0.2	93.24	Learning rate = 0.001 Correct rate = 0.7 Incorect rate = 0.3	84.53

TABLE II
Multimodal optimization RESULT

method	hyper-parameter	accuracy
Grid Search	Learning rate = 0.0001 Correct rate = 0.8 Incorrect rate = 0.2	89.73
Random Search	Learning rate = 0.0001 Correct rate = 0.8 Incorrect rate = 0.2	**92.67**
Bayesian Optimization	Learning rate = 0.001 Correct rate = 0.8 Incorrect rate = 0.2	91.88

VI. Conclusion

In this paper, we explore an approach to improve the classification performance of multimodal data using Spiking Neural Networks (SNNs) through hyper-parameter optimization. We employ SNNs to simultaneously process both image and audio data, effectively performing classification tasks on this multimodal information. The utilization of hyper-parameter optimization techniques to enhance the classification capabilities of the SNNs model represents a significant advancement in the field of multimodal data classification. These research findings have the potential to be applied in practical scenarios that involve the processing and classification of multimodal data, thereby increasing the applicability of SNN models that mimic biological neural processes.

References

[1] W. Maass, “Networks of spiking neurons: The third generation of neural network models,” Neural Netw., vol. 10, no. 9, pp. 1659–1671, 1997.

[2] Schuman, Catherine D., et al. ”Opportunities for neuromorphic computing algorithms and applications.” Nature Computational Science 2.1 (2022): 10-19.

[3] E. Brochu, V. M. Cora, and N. De Freitas, “A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning,” arXiv preprint arXiv:1012.2599, 2010.

[4] Putra, Rachmad Vidya Wicaksana, and Muhammad Shafique. ”Fspinn: An optimization framework for memory-efficient and energy-efficient spiking neural networks.” IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 39.11 (2020): 3601-3613.

[5] P. Diehl and M. Cook, “Unsupervised learning of digit recognition using spike-timing-dependent plasticity,” Front. Comput. Neurosci., vol. 9, p. 99, Aug. 2015.

[6] Chakraborty, Koustav, Asmita Talele, and Savitha Upadhya. ”Voice recognition using MFCC algorithm.” International Journal of Innovative Research in Advanced Engineering (IJIRAE) 1.10 (2014): 2349-2163.

[7] Ahmed, Nasir, T Natarajan, and Kamisetty R. Rao. ”Discrete cosine transform.” IEEE transactions on Computers 100.1 (1974): 90-93.

[8] Jason K. Eshraghian, Max Ward, Emre Neftci, Xinxin Wang, Gregor Lenz, Girish Dwivedi, Mohammed Bennamoun, Doo Seok Jeong, and Wei D. Lu. “Training Spiking Neural Networks Using Lessons From Deep Learning”. Proceedings of the IEEE, 111(9) September 2023.

[9] Bergstra, James, and Yoshua Bengio. ”Random search for hyper-parameter optimization.” Journal of machine learning research 13.2 (2012).

[10] E. Brochu, V. M. Cora, and N. De Freitas, “A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning,” arXiv preprint arXiv:1012.2599, 2010.