SEQUENCE APPROXIMATION USING FEEDFORWARD SPIKING NEURAL NETWORK FOR SPATIOTEMPORAL LEARNING: THEORY AND OPTIMIZATION METHODS

Anonymous authors
Paper under double-blind review

ABSTRACT

A dynamical system of spiking neurons with only feedforward connections can classify spatiotemporal patterns without recurrent connections. However, the theoretical construct of a feedforward Spiking Neural Network (SNN) for approximating a temporal sequence remains unclear, making it challenging to optimize SNN architectures for learning complex spatiotemporal patterns. In this work, we establish a theoretical framework to understand and improve sequence approximation using a feedforward SNN. Our framework shows that a feedforward SNN with one neuron per layer and skip-layer connections can approximate the mapping function between any arbitrary pairs of input and output spike train on a compact domain. Moreover, we prove that heterogeneous neurons with varying dynamics and skip-layer connections improve sequence approximation using feedforward SNN. Consequently, we propose SNN architectures incorporating the preceding constructs that are trained using supervised backpropagation-through-time (BPTT) and unsupervised spiking-timing-dependent plasticity (STDP) algorithms for classification of spatiotemporal data. A Dual Search-space Bayseian Optimization method is developed to optimize architecture and parameters of the proposed SNN with heterogeneous neuron dynamics and skip-layer connections.

1 INTRODUCTION

Spiking neural network (SNN) (Ponulak & Kasinski, 2011) uses biologically inspired neurons and synaptic connections trainable with either biological learning rules such as spike-timing-dependent plasticity (STDP) (Gerstner & Kistler, 2002) or statistical training algorithms such as backpropagation-through-time (BPTT) (Werbos, 1990). The SNNs with simple leaky integrate-and-fire (LIF) neurons and supervised training have shown classification performance similar to deep neural networks (DNN) while being energy efficient (Kim et al., 2020b; Wu et al., 2019; Srinivasan & Roy, 2019). One of SNN's main difference from DNN is that the neurons are dynamical systems with internal states evolving over time, making it possible for SNN to learn temporal patterns without recurrent connections. Empirical results on feedforward-only SNN models show good performance for spatiotemporal data classification, using either supervised training (Lee et al., 2016; Kaiser et al., 2020; Khoei et al., 2020), or unsupervised learning (She et al., 2021). However, while empirical results are promising, a lack of theoretical understanding of sequence approximation using SNN makes it challenging to optimize performance on complex spatiotemporal datasets.

In this work, we develop a theoretical framework for analyzing and improving sequence approximation using feedforward SNN. We view a feedforward connections of spiking neurons as a spike propagation path, hereafter referred to as a *memory pathways* (She et al., 2021), that maps an input spike train with an arbitrary frequency to an output spike train with a target frequency. Consequently, we argue that an SNN with many memory pathways can approximate a temporal sequence of spike trains with time-varying unknown frequencies using a series of pre-defined output spike trains with known frequencies. Our theoretical framework aims to first establish SNN's ability to map frequencies of input/output spike trains within arbitrarily small error; and next, derive the basic principles for adapting neuron dynamics and SNN architecture to improve sequence approximation. The theoretical derivations are then investigated with experimental studies on feedforward SNN for spatiotemporal classifications. We adopt the basic design principles for improving sequence approx-

imation to optimize SNN architectures and study whether these networks can be trained to improve performance for spatiotemporal classification tasks. The key contributions of this work are:

- We prove that any spike-sequence-to-spike-sequence mapping functions on a compact domain can be approximated by feedforward SNN with one neuron per layer using skip-layer connections, which cannot be achieved if no skip-layer connection is used.
- We prove that using heterogeneous neurons having different dynamics and skip-layer connection increases the number of memory pathways a feedforward SNN can achieve and hence, improves SNN's capability to represent arbitrary sequences.
- We develop complex SNN architectures using the preceding theoretical observations and experimentally demonstrate that they can trained with supervised BPTT and unsupervised STDP for classification on spatiotemporal data.
- We develop a dual-search-space Bayesian optimization process to optimize network architecture, neuron dynamics, and hyper-parameters of a feedforward SNN considering heterogeneity and skip-layer connection to improve learning and classification of spatiotemporal patterns.

We experimentally demonstrate that our network design principles coupled with the dual-search-space Bayesian optimization improve classification performance on DVS Gesture (Amir et al., 2017), N-caltech (Orchard et al., 2015), and sequential MNIST. Results show that the design principles derived using our theoretical framework for sequence approximation can improve spatiotemporal classification performance of SNN.

2 Related Work

Most theoretical approaches to analyze SNN (Amit & Huang, 2010; Brea et al., 2013) focus on the storage and retrieval of precise spike patterns, which is different from the approximation of spike-sequence-to-spike-sequence mappings functions. SNN that incorporates excitatory and inhibitory signal is shown for its ability to emulate sigmoidal networks (Maass, 1997) and is theoretically capable of universal function approximation. Feedforward SNN with specially designed spiking neuron models (Iannella & Back, 2001; Torikai et al., 2008) have been demonstrated for function approximation, while for networks using LIF neurons, function approximation has been shown with only empirical results (Farsa et al., 2015). On the other hand, the existing works that has developed efficient training process for SNN and demonstrated classification performance comparable to deep learning models, have mostly used simpler and generic LIF neuron models (Lee et al., 2016; Kaiser et al., 2020; Kim et al., 2020b; Wu et al., 2019; Sengupta et al., 2019; Safa et al., 2021; Han et al., 2020). Therefore, this paper develops the theoretical basis for function approximation using feedforward SNN with LIF neurons, and studies applications of the developed theoretical constructs in improving SNN-based spatiotemporal pattern classification.

The effectiveness of heterogeneous neurons (She et al., 2021) and skip-layer connections (Srinivasan & Roy, 2019; Sengupta et al., 2019) in SNN has been empirically studied in the past. However, no theoretical approach has been presented to understand why such methods improve learning of spike sequences, and how to optimize a SNN's architecture and parameters to effectively exploit these design constructs. It is possible to search for the optimal SNN configurations through optimization algorithms, but the large number of hyper-parameters for spiking neurons and network structure creates a high-dimensional search space that is long and difficult to solve. Bayesian optimization (Snoek et al., 2012) uses collected data points to make decisions on the next test point that could provide improvement, thus accelerates the optimization process. Prior works (Parsa et al., 2019; Kim et al., 2020a) have shown that SNN performance can also be effectively improved with Bayesian optimization. While those works consider a single or a few neuron parameters, the dual-search-space Bayesian optimization proposed in this work optimizes both network architecture and neuron parameters efficiently by separating the discrete search spaces from the continuous search spaces.

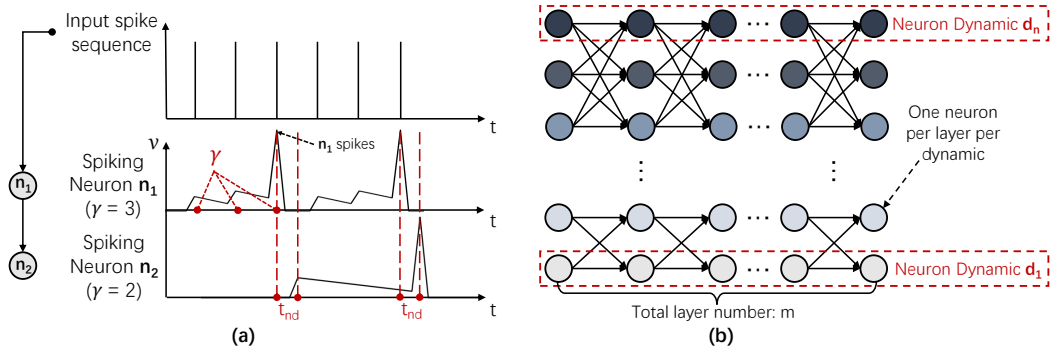

Figure 1: (a) Receiving the given input spike sequence, neuron n_1 needs to receive 3 input spikes to reach threshold thus has neuron response rate $\gamma = 3$. Neuron n_2 receives spike from neuron n_1 with t_{nd} delay, and has $\gamma = 2$. (b) A minimal multi-neuron-dynamic (mMND) network with m layers and n different neuron dynamics.

3 Approximation Theory of Feedforward SNN

3.1 Definitions and Notations

Definition 1 *Neuron Response Rate* γ For a spiking neuron n with membrane potential at v_{reset} and input spike sequence with period t_{in}, γ is the number of input spike n needs to reach v_{th}.

Definition 2 *Neuron Delay* t_{nd} The time for a spike from pre-synaptic neuron to arrive at its post-synaptic neurons.

Definition 3 *Minimal-layer-size Network* A minimal-layer-size network is a feedforward spiking neural network with a finite number of layers and one neuron in each layer.

Definition 4 *Memory Pathways* For a feedforward SNN with m layers, a memory pathway is defined as a spike propagation path connected by neurons in m-tuple $\mathbb{P} = \{D_1, D_2, D_3, ..., D_m\}$ where D_i is the set of neurons included in layer i. $\mathbb{P}$ and $\mathbb{P}'$ are considered to be distinct if

$$\forall\ D_i \in \mathbb{P} \text{ and } D_i' \in \mathbb{P}',\ \exists\ i\ s.t.\ D_i \neq D_i'$$

Definition 5 *Skip-layer Connection* For a feedforward SNN with m layers, a skip-layer connection is defined with source layer and target layer pair (l_s, l_t), such that $l_s \in \{1, 2, 3, ..., (m-2)\}$, and $l_t \in \{(l_s+2), (l_s+3), (l_s+4), ..., m\}$. The output feature map from source layer is concatenated to the original input feature map of the target layer.

Definition 6 *Minimal Multi-neuron-dynamic (mMND) Network* A densely connected network in which each layer has an arbitrary number of neurons that have different neuron parameters. All synapses from one pre-synaptic neuron have the same synaptic conductance.

For a minimal-layer-size network with two layers as shown in Figure 1(a) receiving an input spike sequence with certain period t_{in}, the two neurons in the network have $\gamma = 2$ and $\gamma = 3$, respectively. An example of mMND network with m layers and n neuron dynamics is shown in Figure 1(b). SNN with multilayer perceptron (MLP) structure can be considered a scaled-up mMND network with multiple neurons for each dynamic. A network with convolutional structure can be considered a scaled-up mMND network with duplicated connections in each layer. We analyze the correlation of network capacity and structure based on mMND networks, as for MLP-SNN and Conv-SNN network the analysis can be extended according to the specific layer dimensions.

Notations For the analysis of spike sequence in temporal space, the notation of T_{max} and T_{min} are defined as positive real numbers such that $T_{max} > T_{min}$. $\epsilon > 0$ is the error of approximation.

3.2 MODELING OF SPIKING NEURON

SNN consists of spiking neurons connected with synapses. The spiking neuron model studied in this work is leaky integrate-and-fire (LIF) as defined by the following equations:

$$\tau_m \frac{dv}{dt} = a + R_m I - v;\ v = v_{reset},\ \text{if } v > v_{threshold} \tag{1}$$

R_m is membrane resistance, $\tau_m = R_m C_m$ is time constant and C_m is membrane capacitance. a is a parameter used to adjust neuron behavior during simulation. I is the sum of current from all input synapses that connect to the neuron. A spike is generated when membrane potential v cross threshold and the neuron enters refractory period r, during which the neuron maintains its membrane potential at v_{reset}. The time it take for a pre-synaptic neuron to send a spike to its post-synaptic neurons is t_{nd}. Neuron response rate γ is a property of a spiking neuron's response to certain input spike sequence. We show how the value of γ can be evaluated below.

Remark For any input spike sequence, individual spike can be described with Dirac delta function $\delta(t - t^i)$ where t^i is the time of the i-th input spike. For the membrane potential of a spiking neuron receiving the input before reaching spiking threshold, with initial state at $t = 0$ with $v = v_{reset}$, solving the differential equation (1) leads to:

$$v(t) = v_{reset} e^{-\frac{t}{\tau_m}} + a(1 - e^{-\frac{t}{\tau_m}}) + \frac{R_m}{\tau_m} e^{-\frac{t}{\tau_m}} \sum_i G \int_0^t \delta(t - t_n^i) e^{\frac{t}{\tau_m}} dt \tag{2}$$

Here, G is the conductance of input synapses connected to the neuron. From (2), there exists a timestep u such that $v_m(t^{(u-1)}) < v_{threshold}$ and $v_m(t^u) >= v_{threshold}$. By evaluating (2) for u given neuron parameters and input spike sequence, the neuron response rate γ can be found.

3.3 APPROXIMATION THEOREM OF FEEDFORWARD SNN

To develop the approximation theorem for feedforward SNN, we first aim to understand the range of neuron response rate that can be achieved. We show with Lemma 1 that for any input spike sequence with periods in a closed interval, it is possible to set the neuron response rate γ to any positive integer. Based on this property, we show with Theorem 1 that by connecting a list of spiking neurons with certain γ sequentially and inserting skip-layer connections, approximation of spike-sequence mapping functions can be achieved. To understand whether this capability of feedforward SNN relies on skip-layer connections, we develop Lemma 2 to prove that skip-layer connections are indeed necessary. In subsection 3.4 we investigate the correlation between approximation capability and network structures by analyzing the cutoff property of spiking neurons, which can change the network's connectivity. In our analysis, we focus on two particular designs: heterogeneous network (Lemma 4) and skip-layer connection (Lemma 5), and show their impact on the number of distinct memory pathways in a network. All lemmas are formally proved in the appendix.

Lemma 1 *For any input spike sequence with period t_{in} in range $[T_{min}, T_{max}]$, there exist a spiking neuron n with fixed parameters v_{th}, v_{reset}, a, R_m and τ_m, such that by changing synaptic conductance G, it is possible to set the neuron response rate γ_n to be any positive integer.*

Proof Sketch. Given an input spike sequence, we can derive the highest possible amount of membrane potential Δv within an input period as a function of neuron parameters. We show that it is possible to make Δv tends to zero by configuring the neuron parameters, so that the number of input spikes required to reach threshold can be set to any positive integer by changing G.

Theorem 1 *For any input and target output spike sequence pair with periods $(t_{in}, t_{out}) \in [T_{min}, T_{max}] \times [T_{min}, T_{max}]$, there exist a minimal-layer-size network with skip-layer connections that has memory pathway with output spike period function $P(t)$ such that $|P(t_{in}) - t_{out}| < \epsilon$.*

Proof Sketch. With skip-layer connections, there can be multiple memory pathways in a minimal-layer-size network as neurons can be either included or skipped through. With this property it is possible to create memory pathways with different delay times for each input spike in a network and

by connecting the output of those memory pathways to a common output, spike sequence of any arbitrary period t_{int} such that $t_{int} <= t_{in}$ can be generated within ϵ. By implementing a spiking neuron with response rate γ larger than 1, it is possible to take input spike sequence with t_{int} and create spike sequence with period $\gamma \cdot t_{int}$. This way it is possible to achieve a network with output spike period $P(t)$ such that $|P(t_{in}) - t_{out}| < \epsilon$.

Lemma 2 *With no skip-layer connection, there does not exist a minimal-layer-size network that has output spike period function $P(t)$ such that for any input and target output spike sequence pair with periods $(t_{in}, t_{out}) \in [T_{min}, T_{max}] \times [T_{min}, T_{max}]$, $|P(t_{in}) - t_{out}| < \epsilon$.*

Proof Sketch. A minimal-layer-size network without skip-layer connection has only one memory pathway, and the network can achieve different output spike period only through changing G or the neuron parameters within for the one memory pathway. For a particular input spike sequence with period t_{in}, we show that there exists two output spike periods $P(t_{in})$ and $P(t_{in})'$, such that $P(t_{in}) - P(t_{in})'$ is a constant value independent of network or neuron configurations. Therefore, for any minimal-layer-size network, there exists t_{out} within the range of $(P(t_{in}), P(t'_{in}))$ such that $|P(t_{in}) - t_{out}| < \epsilon$ does not hold true.

3.4 Network Structure and Memory Pathways

Based on Theorem 1, it is possible to approximate an input/output spike sequence mapping function using a minimal-layer-size network with specific configuration, which can be considered as a memory pathway. Since any continuous bounded function on a compact interval can be approximated to arbitrary accuracy using a piece-wise constant function, and it is possible to use a memory pathway to approximate each of the piece-wise constant function, with increasing number of distinct memory pathways, a feedforward SNN can achieve approximation of continuous functions with less error. In this subsection, we show that two SNN structural designs: heterogeneous network i.e. a network having neurons with different dynamics and adding skip-layer connections, a feedforward SNN has the capability to achieve more distinct memory pathways.

Cutoff Frequency of a Memory Path We first show the correlation of cutoff period and spiking neuron parameters with Lemma 3.

Lemma 3 *A spiking neuron has cutoff period $\omega_c = \tau_m \ln(\frac{v_{reset}-a}{v_{reset}-a+\frac{R_m}{\tau_m}G})$ above which input spike sequence cannot cause the spiking neuron to spike.*

Remark From Lemma 3, it can be observed that the cutoff period ω_c of a neuron can be configured to any positive real number by changing the neuron parameters and synaptic conductance G. Further, with fixed G, ω_c can be configured to any positive real number by changing the neuron parameters. Neurons that are in cutoff change the spike propagation path in a network as they send no output spikes. This creates different memory pathways without changing the connections in a network.

Heterogeneous Network If an mMND network has the same parameters for all neurons in each layer, the majority of the neurons are included in the same memory pathway, leading to the upper bound of number of distinct memory pathways to be limited. With Lemma 4, we show the relationship between the upper bound of the number of distinct memory pathways and the number of different neuron dynamics in an mMND network.

Lemma 4 *For an mMND network with m layers and $\{\lambda_1, \lambda_2, ...\lambda_m\}$ number of different neuron dynamics in each layer, the upper bound of the number of distinct memory pathways is $\prod_{i=1}^{m} \lambda_i$.*

Proof Sketch For an mMND network, it is possible to have neurons with different ω_c in each layer, which creates λ_i number of different neuron activation states for layer i. Across all network layers the highest possible number of different neuron activation states is therefore the product of λ of each layer. Since neurons in cutoff do not propagate spikes, they can be removed from a memory pathway. This leads to $\prod_{i=1}^{m} \lambda_i$ as the upper bound of the number of distinct memory pathways.

Compared to a network with homogeneous neuron parameters, in which the upper bound of number of distinct memory pathways is λ_m, Lemma 4 indicates that heterogeneous network increases the maximum achievable number of distinct memory pathways in a feedforward SNN.

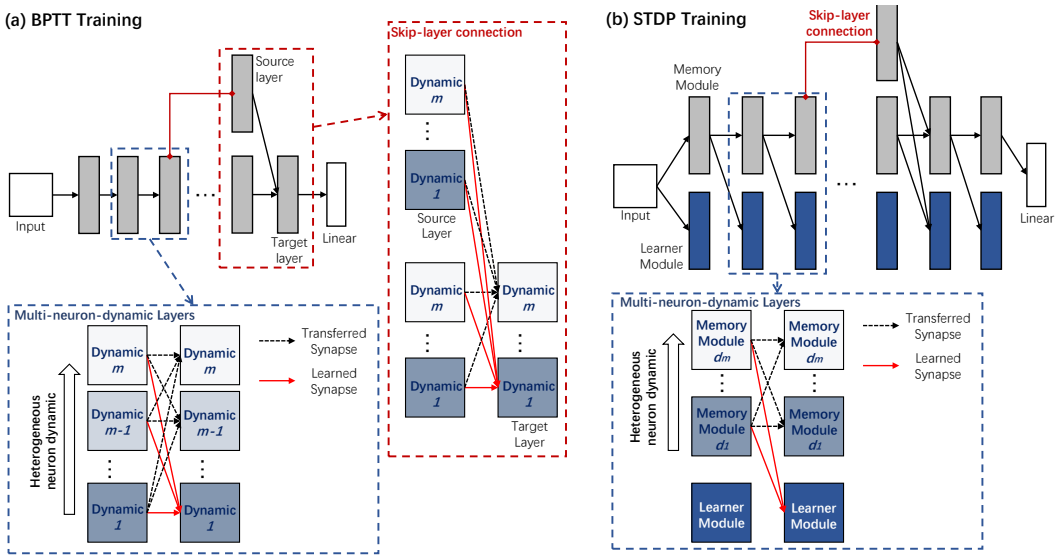

Figure 2: (a) The proposed network with BPTT training, each multi-neuron-dynamic layer contains a set of neuron dynamics from d_1 to d_m. (b) The proposed network with STDP training.

Skip-layer Connection We show that adding skip-layer connection increases the upper bound of the number of memory pathways in a network with Lemma 5.

Lemma 5 *For an mMND network with m layers and $\{\lambda_1, \lambda_2, ...\lambda_m\}$ different neuron dynamics in each layer, a skip-layer connection made between layer l_a and l_b, s.t. $a, b \in \{1, 2, ...m\}$ and $(b-a) > 1$ increases the upper bound of the number of distinct memory pathways to $\prod_{i=1}^{m} \lambda_i + (\prod_{i=1}^{a} \lambda_i \cdot \prod_{i=b}^{m} \lambda_i)$*

Proof Sketch The main idea is similar to the proof of Lemma 4. By adding skip-layer connection, there are additional possible neuron activation states in the network that result from the cutoff of neurons in layers between l_a and l_b. Without layers between l_a and l_b in the spike propagation path, the number of achievable memory pathways is increased by the upper bound of number of distinct memory pathways in layers before l_a and after l_b.

4 SNN Architecture Design using Approximation Theory

In this section, we discuss design of SNN architectures as inspired by the developed approximation theory for feedforward SNN.

Network Template for BPTT Training For BPTT training, the network template is shown in Figure 2(a). Each multi-neuron-dynamic layer, which can either be convolutional or fully connected, uses different neuron parameters for each feature map. There are two types of synapses between layers: transferred synapses marked as black dashed arrows and learned synapses marked as red solid arrows. The conductance of learned synapses is optimized by the BPTT algorithm during training, and the transferred synapses have the same conductance as the learned synapses from the same pre-synaptic neuron. For example, the synapses connecting neurons with dynamic d_m to neurons with dynamic $\{d_2, d_3, d_4, ..., d_{m-1}\}$ in the next layer have conductance transferred from synapses connecting neurons with dynamic d_m to neurons with dynamic d_1. The skip-layer connection is implemented with the output spike matrix from source layer concatenated to the original input spike matrix of the target layer. The skip-layer connection has the same implementation as the regular connection between consecutive layers, with both learned and transferred synapses (Figure 2(a)). The last layer the network is a full-connected layer with homogeneous dynamic to generate prediction labels.

Network Template for STDP Learning For networks trained with STDP, the template is shown in Figure 2(b). Each layer contains a learner module and a memory module. Learner modules use

homogeneous neuron dynamic that is suitable for STDP learning, and memory modules consist of neurons with different dynamics. Similar to BPTT training, there are two types of synapses: transferred synapses and learned synapses. Between two layers, memory modules are connected with transferred synapses and memory modules are connected to learner modules with learned synapses. Leaner modules between layers are not directly connected. STDP training proceeds as a layer-by-layer process. During training of the first layer, conductance of synapses connecting neurons in memory module to neurons in learner modules, referred to as learned synapses, is learned with STDP using all training data without labels. Then, the learned conductance is transferred to the corresponding transferred synapses. The memory module is then used to perceive input patterns and generate spikes during training of the next layer. This lay-by-layer process is repeated until the layer before the final layer finishes learning. The final linear layer is then fine-tuned using stochastic gradient descent (SGD) based on spike frequency array from the last multi-neuron-dynamic layer generated based on the labeled data. Skip-layer connection is implemented by connecting the memory module of the source layer to the target layer. The connections are made with the two types of synapses and follow the same training process as the consecutive layers.

Dual-search-space Bayesian Optimization Bayesian optimization uses Gaussian process to model the distribution of an objective function, and an acquisition function to decide points to evaluate. Formally, the problem in this work is defined as: for unknown function $f : \mathrm{X} \to \mathbb{R}$ that maps network configuration $x \in \mathrm{X}$ to validation accuracy for a certain dataset, find: $x^* \in \mathrm{argmax}_{x \in \mathrm{X}} f(x)$. Since the configuration of heterogeneous network and skip-layer connection have discrete values for their configuration while neuron parameters have continuous values, we implement a dual-search-space optimization process, where the network structural design is first optimized with fixed, manually selected neuron parameters. After an optimal structure is found, neuron parameters are optimized for the selected network structure. This separates the search space of network structure, which is discrete, from the continuous search space of neuron parameters to reduce time consumption for the Bayesian optimization process. In addition, we further improve optimization efficiency by implementing constraints on the search spaces. Details on the configurations of the optimization process are listed in the appendix.

To achieve Bayesian optimization with constraints, we implement a modified expected improvement (EI) acquisition function similar to the one shown by Gardner (Gardner et al., 2014), which uses a Gaussian process to model the feasibility indicator due to its high evaluation cost. In this work, since the constraint function can be explicitly defined, we use feasibility indicator that is directly evaluated. The modified EI function is defined as: $I_c(\mathbf{x}) = \Delta(\mathbf{x}) \cdot \max\{0, P(\mathbf{x}) - P(\mathbf{x}^+)\}$, where $P(\mathbf{x})$ is the objective function to maximize. ($\mathbf{x}^+$ is the point that provided the highest objective function value among all tested points. $\Delta(\mathbf{x})$ is the explicitly defined indicator function that takes the value of 1 when all constraints are satisfied and 0 otherwise.

5 Experiments

5.1 Experiment Settings

Datasets tested in the experiment include the DVS Gesture dataset (Amir et al., 2017), which is an event-based human gesture classification dataset captured by DVS cameras, and the N-Caltech101 (Orchard et al., 2015), which is an event-based version of the Caltech101 dataset. The proposed method is also tested for MLP-style SNN on the sequential MNIST dataset, in which the original MNIST images are presented row-by-row sequentially. We also vary the amount of *labeled data used during training* ranging from using 100% labeled data for training down to 10% labeled data (30% for N-Caltech101) during training. Note, during STDP training networks always uses the entire but *un-labeled* training dataset; however, only the fraction of the labeled data is used for supervised fine-tuning of the last layer. Comparison is made for DVS Gesture and N-Caltech101 with prior works including ConvLSNN, which is a combination of convolutional SNN and recurrent SNN with long and short-term neurons trained with BPTT (Salaj et al., 2020), DECOLLE (Kaiser et al., 2020), which uses surrogate gradient to train a convolutional feedforward SNN, HATS (Sironi et al., 2018), which implements time surfaces and SVM for classification and H-SNN (She et al., 2021) which uses STDP to train a convolutional SNN with two neuron dynamics.

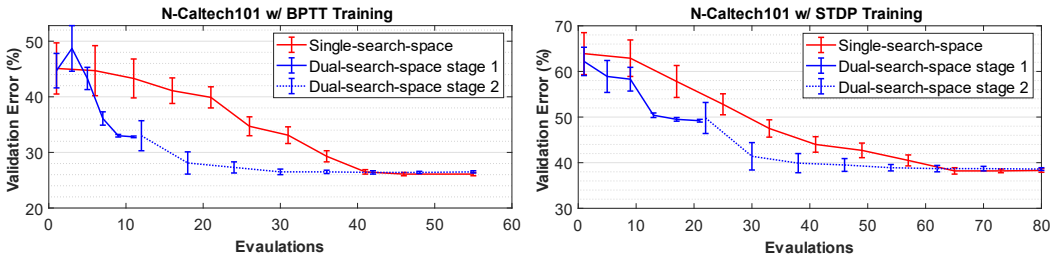

Figure 3: Validation error over optimization iterations for the proposed dual-search-space Bayesian optimization compared to the normal single-search-space Bayesian optimization.

5.2 Effect of Dual-search-space Bayesian Optimization

We compare the proposed dual-search-space Bayesian optimization with regular Bayesian optimization using a single search space for network validation error over 5 runs. The result from the N-Caltech101 dataset is shown in Figure 3. It can be observed that the two optimization approaches achieve similar minimum validation error after convergence. By separating the search spaces, the proposed optimization process reaches convergence faster than regular single-search-space optimization. It is also worth noting that, between the two stages in the optimization process for BPTT training, the first stage accounts for more reduction in validation error than the second stage. This indicates that optimizing network structure causes more impact to BPTT training than optimizing neuron parameters, which is potentially due to the reason that network structures more heavily affects the number of memory pathways in the network than neuron parameters. On the other hand, for STDP training where learning behavior is sensitive to the dynamic of spiking neurons, the reduction of validation error is more equally shared between the two optimization stages. Over the 5 runs, among all network configurations achieved after the dual-search-space optimization converges, we compare the configuration with the lowest number of trainable parameters against baseline models. The specific configurations for the optimized networks are listed in Table 1. It can be observed that for BPTT algorithm, the optimized networks have more layers than the STDP trained networks, and the optimal values found for neuron parameters are highly distinct for the two training methods.

Table 1: Configuration of optimized network models

Network	Conv. Layer Number	Skip-layer Connection	Number of Different Neuron Dynamics and a	Neuron Parameters τ_m	Neuron Parameters R_m
BPTT, Gesture	9	(2,7)	4, (-24,-17,-12,-9)	120	340
BPTT, N-Caltech	12	(2,5), (5,8), (8,11)	5, (-23,-16,-14,-11,-8)	70	300
STDP, Gesture	6	(2,4), (4,6)	4, (-26,-24,-15,-9)	110	260
STDP, N-Caltech	8	(3,5), (5,7)	6, (-21,-19,-17,-13,-9,-7)	140	240

5.3 Ablation Studies

To investigate the effect of using multiple neuron dynamics, we apply the same dual-search-space Bayesian optimization process for networks that have homogeneous neuron dynamic for the same

Table 2: Ablation studies of optimized networks

Model	Heterogeneity	Skip-layer	DVS Gesture	N-Caltech101	S-MNIST
Homogeneous-BPTT	N	Y	95.0	65.3	95.5
No-skip-layer-BPTT	Y	N	96.5	63.5	94.8
This Work-BPTT	Y	Y	98.0	71.2	97.3
Homogeneous-STDP	N	Y	91.3	37.0	94.3
No-skip-layer-STDP	Y	N	93.1	51.9	95.5
This Work-STDP	Y	Y	96.6	58.1	96.1

Table 3: Accuracy (%) for DVS Gesture (top) and N-Caltech101 (bottom)

Model	Labeled Data % In Training 100%	50%	30%	10%	Parameter No.
ConvLSNN (Salaj et al., 2020)	97.1	95.3	92.0	84.3	2.9M
DECOLLE (Kaiser et al., 2020)	97.5	95.0	91.2	83.9	1.3M
HATS (Sironi et al., 2018)	95.2	94.1	91.6	83.7	-
H-SNN (She et al., 2021)	96.2	95.8	93.7	88.2	0.74M
This Work-STDP Training	96.6	**96.0**	**94.1**	**91.2**	0.81M
This Work-BPTT Training	**98.0**	95.3	91.1	82.4	1.1M

Model	Labeled Data % In Training 100%	70%	50%	30%	Parameter No.
ConvLSNN (Salaj et al., 2020)	63.1	58.7	51.3	45.4	3.0M
DECOLLE (Kaiser et al., 2020)	66.9	61.9	56.2	50.6	2.0M
HATS (Sironi et al., 2018)	64.2	61.0	54.3	48.8	-
H-SNN (She et al., 2021)	42.8	41.9	37.0	34.6	1.7M
This Work-STDP Training	58.1	57.8	**57.2**	**54.6**	1.4M
This Work-BPTT Training	**71.2**	**65.4**	56.0	52.5	1.7M

number of evaluations as the proposed design. Similarly, to study the contribution to performance gain from skip-layer connections, the Bayesian optimization process is used for network templates without skip-layer connections. The optimization process runs for the same number of evaluations as the proposed design. From the results shown in Table 2, it can be observed that compared to baselines, the proposed networks achieve the best accuracy for all datasets. Specifically, when homogeneous network is used, the performance of STDP trained network is noticeably lower than the proposed method for DVS Gesture and N-Caltech101. For BPTT training, using heterogeneous network and skip-layer connection shows different level of benefit for each dataset. For sequential MNIST which has less complexity, the improvement from using heterogeneous neurons and skip-layer connections is not as significant.

5.4 Comparison with Prior Works

DVS Gesture With 100% labels available during training, the proposed network trained with BPTT demonstrates higher accuracy than all tested networks with the less trainable parameters than all baselines in Table 3 (top). The proposed network trained with STDP has slightly lower accuracy than ConvLSNN and DECOLLE when 100% labels are used; for reduced-label training it outperforms all network (including H-SNN).

N-Caltech101 As shown in Table 3 (bottom), the proposed network trained with BPTT outperforms all baselines with both 70% and 100% training labels and also has less number of trainable parameters. The un-supervised learning models i.e., H-SNN and the proposed network with STDP, show considerably lower performance (more than what was observed for DVS Gesture) than supervised ones when 100% labels are available; However, the proposed network with STDP shows better performance than H-SNN, and outperforms all network when available labels are below 50%.

6 Conclusion

We develop a theoretical basis to understand and optimize the ability of a feedforward SNN to approximate a temporal sequence. We analytically show how heterogeneity and skip-layer connections can improve sequence approximation with SNN, and empirically demonstrate their impact on spatiotemporal learning. It is well-known in neuroscience that, heterogeneity (De Kloet & Reul, 1987) and irregular connectivity (Eickhoff et al., 2018) are intrinsic properties of human brains. Our analysis shows that incorporating such concepts within artificial SNN is beneficial for designing high-performance SNN for classification of spatiotemporal data.

REFERENCES

Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7243–7252, 2017.

Y. Amit and Yibi Huang. Precise capacity analysis in binary networks with multiple coding level inputs. *Neural Computation*, 22:660–688, 2010.

Johanni Brea, Walter Senn, and Jean-Pascal Pfister. Matching recall and storage in sequence learning with spiking neural networks. *Journal of neuroscience*, 33(23):9565–9575, 2013.

ER De Kloet and JMHM Reul. Feedback action and tonic influence of corticosteroids on brain function: a concept arising from the heterogeneity of brain receptor systems. *Psychoneuroendocrinology*, 12(2):83–105, 1987.

Simon B Eickhoff, BT Thomas Yeo, and Sarah Genon. Imaging-based parcellations of the human brain. *Nature Reviews Neuroscience*, 19(11):672–686, 2018.

Edris Zaman Farsa, Soheila Nazari, and Morteza Gholami. Function approximation by hardware spiking neural network. *Journal of Computational Electronics*, 14(3):707–716, 2015.

Jacob R Gardner, Matt J Kusner, Zhixiang Eddie Xu, Kilian Q Weinberger, and John P Cunningham. Bayesian optimization with inequality constraints. In *ICML*, volume 2014, pp. 937–945, 2014.

Wulfram Gerstner and Werner M Kistler. Mathematical formulations of hebbian learning. *Biological cybernetics*, 87(5-6):404–415, 2002.

Bing Han, Gopalakrishnan Srinivasan, and Kaushik Roy. Rmp-snn: Residual membrane potential neuron for enabling deeper high-accuracy and low-latency spiking neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Nicolangelo Iannella and Andrew D. Back. A spiking neural network architecture for nonlinear function approximation. *Neural Networks*, 14(6):933–939, 2001. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(01)00080-6. URL https://www.sciencedirect.com/science/article/pii/S0893608001000806.

Jacques Kaiser, Hesham Mostafa, and Emre Neftci. Synaptic plasticity dynamics for deep continuous local learning (decolle). *Frontiers in Neuroscience*, 14:424, 2020. ISSN 1662-453X. doi: 10.3389/fnins.2020.00424. URL https://www.frontiersin.org/article/10.3389/fnins.2020.00424.

Mina A Khoei, Amirreza Yousefzadeh, Arash Pourtaherian, Orlando Moreira, and Jonathan Tapson. Sparnet: Sparse asynchronous neural network execution for energy efficient inference. In *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pp. 256–260. IEEE, 2020.

Seijoon Kim, Seongsik Park, Byunggook Na, Jongwan Kim, and Sungroh Yoon. Towards fast and accurate object detection in bio-inspired spiking neural networks through bayesian optimization. *IEEE Access*, 9:2633–2643, 2020a.

Seijoon Kim, Seongsik Park, Byunggook Na, and Sungroh Yoon. Spiking-yolo: Spiking neural network for energy-efficient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11270–11277, 2020b.

Jun Haeng Lee, Tobi Delbruck, and Michael Pfeiffer. Training deep spiking neural networks using backpropagation. *Frontiers in Neuroscience*, 10:508, 2016. ISSN 1662-453X. doi: 10.3389/fnins.2016.00508. URL https://www.frontiersin.org/article/10.3389/fnins.2016.00508.

Wolfgang Maass. Fast sigmoidal networks via spiking neurons. *Neural computation*, 9(2):279–304, 1997.

Garrick Orchard, Ajinkya Jayawant, Gregory K. Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in Neuroscience*, 9:437, 2015. ISSN 1662-453X. doi: 10.3389/fnins.2015.00437. URL https://www.frontiersin.org/article/10.3389/fnins.2015.00437.

Maryam Parsa, J Parker Mitchell, Catherine D Schuman, Robert M Patton, Thomas E Potok, and Kaushik Roy. Bayesian-based hyperparameter optimization for spiking neuromorphic systems. In *2019 IEEE International Conference on Big Data (Big Data)*, pp. 4472–4478. IEEE, 2019.

Filip Ponulak and Andrzej Kasinski. Introduction to spiking neural networks: Information processing, learning and applications. *Acta neurobiologiae experimentalis*, 71(4):409–433, 2011.

Ali Safa, Federico Corradi, Lars Keuninckx, Ilja Ocket, André Bourdoux, Francky Catthoor, and Georges GE Gielen. Improving the accuracy of spiking neural networks for radar gesture recognition through preprocessing. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Darjan Salaj, Anand Subramoney, Ceca Kraišniković, Guillaume Bellec, Robert Legenstein, and Wolfgang Maass. Spike-frequency adaptation provides a long short-term memory to networks of spiking neurons. *bioRxiv*, 2020. doi: 10.1101/2020.05.11.081513. URL https://www.biorxiv.org/content/early/2020/05/12/2020.05.11.081513.

Abhronil Sengupta, Yuting Ye, Robert Wang, Chiao Liu, and Kaushik Roy. Going deeper in spiking neural networks: Vgg and residual architectures. *Frontiers in neuroscience*, 13:95, 2019.

Xueyuan She, Saurabh Dash, Daehyun Kim, and Saibal Mukhopadhyay. A heterogeneous spiking neural network for unsupervised learning of spatiotemporal patterns. *Frontiers in Neuroscience*, 14:1406, 2021.

A. Sironi, M. Brambilla, N. Bourdis, X. Lagorce, and R. Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1731–1740, 2018. doi: 10.1109/CVPR.2018.00186.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

Gopalakrishnan Srinivasan and Kaushik Roy. Restocnet: Residual stochastic binary convolutional spiking neural network for memory-efficient neuromorphic computing. *Frontiers in neuroscience*, 13:189, 2019.

Hiroyuki Torikai, Atsuo Funew, and Toshimichi Saito. Digital spiking neuron and its learning for approximation of various spike-trains. *Neural Networks*, 21(2):140–149, 2008. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2007.12.045. URL https://www.sciencedirect.com/science/article/pii/S0893608008000051. Advances in Neural Networks Research: IJCNN '07.

Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, Yuan Xie, and Luping Shi. Direct training for spiking neural networks: Faster, larger, better. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1311–1318, 2019.

A SNN DYNAMICS

Remark For a sequentially connected neuron list with m neurons all with $\gamma = 1$ and neuron delay t_{nd}, an input spike at time t leads the neuron list to generate an output spike at time $t + mt_{nd}$

Remark For any input sequence with period t_{in} to a spiking neuron with response rate γ such that $\gamma > 1$, if refractory period is set to $r < t_{in}$, the neuron can exit refractory period before the next spike arrives.

Lemma 1 For any input spike sequence with period t_{in} in range $[T_{min}, T_{max}]$, there exist a spiking neuron n with fixed parameters v_{th}, v_{reset}, a, R_m and τ_m, such that by changing synaptic conductance G, it is possible to set the neuron response rate γ_n to be any positive integer.

Proof. For a given input spike sequence period t_{in}, we consider the maximum value of membrane potential decay that can happen. From (1), without input spike, $\frac{dv}{dt}$ is negative and it's absolute value increases with higher v. We therefore consider the decay from v such that $v \to v_{th}^-$ for period T_{max}. The value of $v(t)$ at $t = T_{max}$ can be found by solving the differential equation (1) for $v(t) = v_{th}$ at $t = 0$:

$$v(T_{max}) = v_{th}e^{-\frac{T_{max}}{\tau_m}} - ae^{-\frac{T_{max}}{\tau_m}} + a$$

It is possible to have a spiking neuron with R_m, a and τ_m such that Δv, defined as

$$\Delta v = v(T_{max}) - v(0) = v_{th}e^{-\frac{T_{max}}{\tau_m}} - ae^{-\frac{T_{max}}{\tau_m}} + a - v_{th}$$

tends to zero. Since the the highest possible decrease of membrane potential is negligible, for any target γ, it is possible to set G such that $G = \frac{v_{th} - v_{reset}}{\gamma}$. The proof is complete.

B PROOF OF THEOREM 1

Theorem 1 For any input and target output spike sequence pair with periods $(t_{in}, t_{out}) \in [T_{min}, T_{max}] \times [T_{min}, T_{max}]$, there exist a minimal-layer-size network with skip-layer connections that has memory pathway with output spike period function $P(t_{in})$ such that $|P(t_{in}) - t_{out}| < \epsilon$.

Proof. For any given (t_{in}, t_{out}), first consider the condition where $t_{in} > t_{out}$. It is possible to construct a minimal-layer-size network N connecting m spiking neurons with neuron response rate $\gamma = 1$ sequentially, denoted as a m-tuple of neurons $\{n_1, n_2, ..., n_m\}$. Since any configuration of skip-layer connection with source layer and target layer pair (l_s, l_t), such that $l_s \in [1, m-2]$, and $l_t \in [l_s + 2, m]$, can be added, it is possible to add a $(m-2)$-tuple of skip-layer connections $S_{sl} = \{(i, m) \ \forall \ i \in \{1, 2, 3, ..., m-2\}\}$. Denote the synaptic conductance for all the skip-layer connections as a $(m-2)$-tuple $S_{G^{sl}} = \{G_1^{sl}, G_2^{sl}, G_3^{sl}, ..., G_{m-2}^{sl}\}$

For any $t_{out} < t_{in}$, it is possible to find a k-tuple of synaptic conductance $S'_{G^{sl}} = \{G_i^{sl}, G_{2i}^{sl}, G_{3i}^{sl}, ..., G_{ki}^{sl}\}$ such that $i = \lfloor \frac{t_{out}}{t_{nd}} \rfloor$ and $k = \lfloor \frac{m-2}{i} \rfloor$. Set synaptic conductance in $S_{G^{sl}} \setminus S'_{G^{sl}}$ to 0. Then set the conductance of synapse connecting n_{m-1} and n_m to 0. In such way, The output spikes from network N has period $P(t_{in}) = \lfloor \frac{t_{out}}{t_{nd}} \rfloor \cdot t_{nd}$. For given ϵ, it is possible to choose t_{nd} such that $t_{nd} < 2\epsilon$, therefore satisfying $|P(t_{in}) - t_{out}| < \epsilon$. m can be chosen as $m = \frac{T_{max} - T_{min}}{t_{nd}}$, or equivalently $m = \frac{T_{max} - T_{min}}{2\epsilon}$. Since $\frac{T_{max} - T_{min}}{2\epsilon}$ is finite, m is finite.

Now we consider $t_{in} < t_{out}$. Using N as described above, it is possible to achieve output spike with period within ϵ of any period in $(0, t_{in}]$. For a given t_{out}, assume the configuration in neuron list N has output spike interval t'_{int} such that $kt'_{int} = t_{out}$, where k is a positive integer. From Lemma 1, it is possible to set G for a neuron n_{m+1} such that, with input spike period t'_{int}, its neuron response delay is $\gamma_{n_{m+1}} = k$. By connecting n_{m+1} to the output of N, the new network, denoted as N', has output spike with period $P(t_{in}) = t'_{int}/k = t_{out}$. Therefore, for a given ϵ, need to have neuron list N with actual output spike interval t_{int} such that,

$$|t_{int} - t'_{int}| < \frac{\epsilon}{k} \tag{3}$$

Since k is finite, (3) can be achieved. Now, for $t_{in} >= t_{out}$, the network N' can be configured such that t_{int} satisfies $|t_{int} - t_{out}| < \epsilon$, and the value of $\gamma_{n_{m+1}}$ set to 1. $|P(t_{in}) - t_{out}| < \epsilon$ can then be achieved. The proof is complete.

C PROOF OF LEMMA 2

Lemma 2 With no skip-layer connection, there does not exist a minimal-layer-size network that has output spike period function $P(t_{in})$ such that for any input and target output spike sequence pair with periods $(t_{in}, t_{out}) \in [T_{min}, T_{max}] \times [T_{min}, T_{max}]$, $|P(t_{in}) - t_{out}| < \epsilon$.

Proof. A minimal-layer-size network N with m layers has a m-tuple of neurons $\{n_1, n_2, ..., n_m\}$ connected sequentially. Since no skip-layer connection is present, there is only one memory pathway, which contains all neurons $\{n_1, n_2, ..., n_m\}$. Denoted the neuron response rate corresponding to each neuron in N as $\Gamma = \{\gamma_1, \gamma_2, ..., \gamma_m\}$. Consider the output spike sequence when $\gamma_i = 1 \, \forall \, \gamma_i \in \Gamma$. For a given input sequence with t_{in} that has the first spike at time t_1, the first output spike has timing of $t_1 + mt_{nd}$, and the second output spike has timing of $t_1 + t_{in} + mt_{nd}$. It can be easily derived that the period of the output spike sequence is t_1. Now consider the output spike sequence when $\gamma_j = 2$ for any $j \in \{1, 2, 3, 4, ..., m\}$ and $\gamma_i = 1 \ \forall \ i \in (\{1, 2, 3, 4, ..., m\} \setminus \{j\})$. Following the same process, the period of the output spike sequence is $\frac{t_1}{2}$. Since the smallest increase to any γ is by 1, there are no set of values for Γ such that the resulting output spike sequence has period between $(t_{in}, 2t_{in})$. Therefore, within the range $(t_{in}, 2t_{in})$, there exists values of t_{out} such that $|P(t_{in}) - t_{out}| < \epsilon$ does not hold. The proof is complete.

D MEMORY PATHWAYS IN SNN

In this section we analyze the increase to the upper bound of the number of memory pathways in a network by using heterogeneous networks and skip-layer connections.

Lemma 3 A spiking neuron has cutoff period $\omega_c = \tau_m \ln(\frac{v_{reset} - a}{v_{reset} - a + \frac{R_m}{\tau_m} G})$ above which input spike sequence cannot cause the spiking neuron to spike.

Proof. Consider (2), since the membrane potential increases at time of t^i and decays otherwise, solving for $t = t^i$ and the equation can be expanded:

$$v_m(t^i) = v_{reset} e^{\frac{t^i}{\tau_m}} + a(1 - e^{\frac{t^i}{\tau_m}}) + \frac{R_m}{\tau_m} G e^{\frac{t^i - t^1}{\tau_m}} + \frac{R_m}{\tau_m} G e^{\frac{t^i - t^2}{\tau_m}} + ... + \frac{R_m}{\tau_m} G$$

For input with frequency f, $t^{i+1} - t^i = \Delta t = \frac{1}{f}$, subtracting membrane potential values at two consecutive t_i provides:

$$\Delta v_m = v_m(t^{i+1}) - v_m(t^i) = v_{reset}(e^{\frac{t^{i+1}}{\tau_m}} - e^{\frac{t^i}{\tau_m}}) - a(e^{\frac{t^{i+1}}{\tau_m}} - e^{\frac{t^i}{\tau_m}}) + \frac{R_m}{\tau_m} G e^{\frac{t^{i+1} - t^1}{\tau_m}} \tag{4}$$

setting time of first input spike t^1 to zero leads to:

$$\Delta v_m = e^{\frac{t^i}{\tau_m}} ((e^{\frac{\Delta t}{\tau_m}} - 1)(v_{reset} - a) + \frac{R_m}{\tau_m} G e^{\frac{\Delta t}{\tau_m}})$$

As $e^{\frac{t^i}{\tau_m}} > 0$, and the term $((e^{\frac{\Delta t}{\tau_m}}) - 1)(v_{reset} - a) + \frac{R_m}{\tau_m} G e^{\frac{\Delta t}{\tau_m}})$ does not depend on t^i, the polarity of Δv_m does not change with time. v_m is either strictly increasing, staying the same or decreasing with higher t^i. This indicates that, when $\Delta v_m \leq 0$ the post-synaptic neuron can never spike regardless of how many pre-synaptic spike it receives. $\Delta v_m \leq 0$ when input spike period t_{in} satisfies

$$t_{in} \geq \tau_m \ln(\frac{v_{reset} - a}{v_{reset} - a + \frac{R_m}{\tau_m} G})$$

Therefore, the cutoff period of the neuron is $\omega_c = \tau_m \ln(\frac{v_{reset} - a}{v_{reset} - a + \frac{R_m}{\tau_m} G})$. The proof is complete.

In the following proof, we consider cutoff frequency, $f_c = \frac{1}{\omega_c}$ of spiking neurons.

Lemma 4 For an mMND network with m layers and $\{\lambda_1, \lambda_2, ...\lambda_m\}$ number of different neuron dynamics in each layer, the upper bound of the number of memory pathways is $\prod_{i=1}^{m} \lambda_i$.

Proof. For each layer of the mMND network, since each neuron has different parameters and there are λ_i neurons in layer i, according to Lemma 3, there can be at most λ_i different cutoff frequencies among all neurons in layer i. We consider the case where the maximum number of different cutoff frequencies, i.e. λ_i, exists for all layers $i \in \{1, 2, 3...m\}$ in the mMND network.

For layer 1, it is possible to sort cutoff frequencies of all neurons $\{f_c^1, f_c^2, ...f_c^{\lambda_1}\}$ into an ordered list. For simplicity, assume the neuron indices already put the list in ascending order: $f_c^1 < f_c^2 < f_c^3, ...f_c^{\lambda_1 - 1} < f_c^{\lambda_1}$. For a particular input spike sequence to layer 1 with frequency $f_{in}\ s.t.\ f_c^{\mu} < f_{in} < f_c^{\mu+1}$, neurons with indices from 1 to μ receive input sequence above their cutoff frequencies and can be activated.

Therefore, among all input spike sequence with frequency $f_{in} > f_c^1$, there can be a total of λ_1 possible neuron activation states in the first layer. The same property also applies to layer 2, which creates $\lambda_1\lambda_2$ possible neuron activation states across layer 1 and 2. Repeat this to the last layer, the number of different neuron activation states throughout the network is $\prod_{i=1}^{m} \lambda_i$. As inactive neuron does not send information to its post-synaptic neurons, the network output will not be affected if the connections between inactive neurons and their corresponding post-synaptic neurons are removed. This means that, there are equivalently $\prod_{i=1}^{m} \lambda_i$ possible memory pathways with different connectivity in the mMND network. While we consider the condition where all layers have the maximum number of different cutoff frequencies, for any mMND networks with m layers, the number of different cutoff frequencies n^{f_c} in all layers satisfies $n_i^{f_c} \leq \lambda_i\ \forall\ i \in \{1, 2, 3..., m\}$. Therefore, the upper bound of the number of distinct memory pathways is $\prod_{i=1}^{m} \lambda_i$. The proof is complete.

Lemma 5 For a mMND network with m layers and $\{\lambda_1, \lambda_2, ...\lambda_m\}$ different neuron dynamics in each layer, a skip-layer connection made between layer l_a and l_b, s.t. $a, b \in \{1, 2, ...m\}$ and $(b-a) > 1$ increases the upper bound of the number of memory pathways to $\prod_{i=1}^{m} \lambda_i + (\prod_{i=1}^{a} \lambda_i \cdot \prod_{i=b}^{m} \lambda_i)$

Proof. For the mMND network with skip-layer connection between layer l_a and layer l_b, denoted as P', we consider the case where the maximum number of different cutoff frequencies, i.e. λ_i, exists for all layers $i \in \{1, 2, 3...m\}$ in the mMND network, and network output feature vector is non-zero. The set of all neuron activation states in P' that generates non-zero network output feature vector can be partitioned into two subsets denoted as A and B. Set A contains all states where the input frequency to any layer l_i such that $a < i < b$ is below cutoff frequencies of all neurons in layer l_i. B contains all states where all layers receive input frequency higher than cutoff frequency of at least one neuron in each layer.

For all states in A, no spike signal is sent from layer $b-1$ to layer b, thus the network output is not affected if connections between layer l_i and l_{i+1}, such that $i \in \{a, a+1, ...b-1\}$, are removed. The network is equivalent to network P'' that contains layers $\{l_1, l_2, ...l_a, l_b, l_{b+1}, ...l_m\}$ connected sequentially. Similar to the proof for Lemma 4, it can be determined that the number of memory pathways of P'' is $\prod_{i=1}^{a} \lambda_i \cdot \prod_{i=b}^{m} \lambda_i$ for all states in set A. For all states in set B, since the activation of neurons in the source layer of the skip-layer connection is already accounted for when considering layer l_a, the number of memory pathways is the same as network P' without skip-layer connection, which is $\prod_{i=1}^{m} \lambda_i$ according to Lemma 4.

For the set of memory pathways from states in A, denoted as M_A, and the set of memory pathways from states in B, denoted as M_B, it satisfies that $M_A \cap M_B = \emptyset$, since all elements in M_A have $(m - (b - a - 1))$ layers, and all elements in M_B have m layers. Therefore, the number of memory pathways of network P', under all circumstances where the network has non-zero output vector, is $|M_A \cup M_B| = \prod_{i=1}^{m} \lambda_i + (\prod_{i=1}^{a} \lambda_i \cdot \prod_{i=b}^{m} \lambda_i)$. While we consider the specific network in which all layers have the maximum possible number of different cutoff frequencies, for any mMND networks with m layers and skip layer connection between l_a and l_b, the number of different cutoff frequencies n^{f_c} in all layers satisfies $n_i^{f_c} \leq \lambda_i\ \forall\ i \in \{1, 2, 3..., m\}$. Therefore, the upper bound of the number of memory pathways is $\prod_{i=1}^{m} \lambda_i + (\prod_{i=1}^{a} \lambda_i \cdot \prod_{i=b}^{m} \lambda_i)$. The proof is complete.

E DETAILS ON THE BAYESIAN OPTIMIZATION PROCESS

During the first stage of the dual-search-space optimization process, the parameters to optimize include: N_{layer}, L_{start}, L_{end}, N_{skip}, $N_{dynamic}$, all of which are positive integers. Specifically, N_{layer} is the number of convolutional layers. For skip-layer connection, there are three configuration parameters to optimize: starting layer L_{start}, which is the source layer of the first skip-layer connection; ending layer L_{end}, which is the target layer of the last skip-layer connection; skip-layer connection number N_{skip}, which defines how many connections to implement. The source layer of the N_{skip} skip-layer connections are placed evenly between L_{start} and L_{end}, each with skip length of $\lfloor (L_{end} - L_{start})/N_{skip} \rfloor$, in case $((L_{end} - L_{start})/N_{skip}) \neq \lfloor (L_{end} - L_{start})/N_{skip} \rfloor$, the value of L_{end} is reduced to the maximum value that satisfies $((L_{end} - L_{start})/N_{skip}) = \lfloor (L_{end} - L_{start})/N_{skip} \rfloor$. For heterogeneity, the number of different dynamic $N_{dynamic}$ in all layers are optimized jointly. The constraint for the parameters is that, $N_{layer} \in [4, 15]$, $2 \leq L_{start} < (N_{layer} - 1)$, $(L_{start} + 1) < L_{end} \leq N_{layer}$, and $0 \leq N_{skip} \leq (L_{end} - L_{start})/2$ and $N_{dynamic} \in [1, 10]$. The manually configured neuron parameters are, $\tau_m = 100$ and $R_m = 300$ for all neuron dynamics, and $a \in [-30, -5]$ is distributed evenly for each neuron dynamics.

Due to the exponential increase of search space with the number of neuron dynamics in each layer, it is highly inefficient to search for every neuron parameters in each dynamic. In the second stage of the optimization process, we choose to apply Bayesian optimization for the parameter a of each neuron dynamic separately, while τ_m and R_m are optimized jointly with the same values shared by all neuron dynamics. a values are taken to the precision of 10^0, and τ_m and R_m values are taken to the precision of 10^1. The constraints are $a \in [-30, -5]$, $\tau_m \in [50, 200]$ and $R_m \in [200, 400]$. The value of t_{nd} for all networks is set to 1. The parameters of each optimized networks are shown in Table 1. Note, the skip-layer connections are listed as source and target layer pairs.