Attention-based canonical microcircuit model for spiking neural network

Lee, Yelim

Department of Medical Science

Graduate School

Yonsei University

Attention-based canonical microcircuit model for spiking neural network

Advisor Park, Hae-Jeong

A Master's Thesis Submitted
to the Department of Medical Science
and the Committee on Graduate School
of Yonsei University in Partial Fulfillment of the
Requirements for the Degree of
Master of Medical Science

Lee, Yelim

June 2025

Attention-based canonical microcircuit model for spiking neural network

This Certifies that the Master's Thesis of Lee, Yelim is Approved

Committee Chair	Chung, Seung Soo
Committee Member	Park, Hae-Jeong
Committee Member	Chun, Sehun

Department of Medical Science
Graduate School
Yonsei University
June 2025

ACKNOWLEDGEMENTS

Above all, I offer my deepest thanks to God, whose guidance and wisdom have sustained me throughout the course of this degree. I am especially thankful for being led to Him and for the grace to reflect on His Word throughout this journey, which has remained a constant source of strength during times of challenge and uncertainty.

I am deeply thankful to my academic advisor Professor Hae-Jeong Park for providing insightful guidance and unwavering support throughout this research. The encouragement to remain curious about the human brain, along with the instruction in asking meaningful questions and conducting rigorous analyses, has been instrumental in my academic development. I am also thankful for his encouragement during difficult times, which enabled me to persevere to the end.

I also wish to thank the members of our laboratory and Dr. Dongmyeong Lee for their intellectual and emotional support during difficult moments. Their presence lightened the burden of challenges and enriched this journey in countless ways.

Finally, I extend my heartfelt appreciation to my family. Their enduring love, encouragement, and care have been an unfailing source of strength throughout this journey. I am especially grateful for their prayers and for always being my true home.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Attention-based canonical microcircuit model for spiking neural network

Selective attention allows organisms to prioritize goal-relevant stimuli under ambiguity. Inspired by neuromodulatory mechanisms—especially cholinergic modulation—we propose a biologically grounded attention mechanism for spiking neural networks (SNNs). Our model dynamically regulates neuronal excitability by modulating firing thresholds using task-driven attention signals.

The network consists of three hidden layers of two-compartment neurons and receives top-down attention cues that guide focus toward either the front or back digit in overlapped MNIST images. Training was performed in two stages: pretraining on standard MNIST digits without attention, and fine-tuning with attention cues on overlapped digits.

Our results show that the attention model significantly outperformed baseline and control models, reducing reversed prediction errors and restoring clean internal representations of target digits. Attention effects were strongest in the deepest hidden layer (Layer 3), which also showed the most efficient learning and energy optimization. Additional analysis revealed that attention operated via both excitation and disinhibition, and that spatial attention was accurately allocated to task-relevant regions.

This study demonstrates the utility of neuromodulatory attention in SNNs and offers a biologically plausible approach for task-dependent selective processing.

Key words: predictive coding, spiking neural network, attention, multi-compartment neurons, classification

1. Introduction

Selective attention enables organisms to prioritize goal-relevant stimuli while suppressing irrelevant information, particularly under conditions of sensory ambiguity[1]. Extensive research has explored this top-down mechanism, demonstrating that higher-order cognitive processes can modulate neuronal excitability and exert gain control to enhance sensitivity to task-relevant inputs[1–3].

Among the various biological mechanisms underlying attention, neuromodulation plays a particularly important role[4–8]. Neuromodulators influence ion channel behavior, synaptic efficacy, and neurotransmitter release through diverse pathways, thereby regulating neuronal excitability and shaping the nonlinear dynamics of membrane and synaptic responses[4;9–11].

In this study, we implemented an attention mechanism in a spiking neural network (SNN) that regulates neuronal excitability through neuromodulation. We investigated whether this mechanism enables the network to exhibit appropriate selective attention in response to task-relevant cue signals under ambiguous visual input conditions. Spiking Neural Networks have gained traction in both neuroscience and machine learning due to their energy efficiency and biological plausibility. While recent studies have incorporated attention into SNNs—often inspired by transformer-based architectures—few have explicitly modeled top-down selective attention through biologically grounded neuromodulatory processes that directly modulate the excitability of individual neurons[12–15].

We present a hierarchical, multi-layer SNN composed of two-compartment neurons, in which attention is mediated through a slow-acting neuromodulatory signal introduced to the soma compartment. This modulation adjusts the neuron's baseline membrane potential, thereby dynamically controlling its excitability[16;17]. This design enables a biologically plausible form of top-down attention and allows the network to flexibly prioritize relevant features based on contextual information.

The canonical microcircuit model explicitly distinguishes between excitatory and inhibitory neurons and defines a fundamental functional unit in the hierarchical organization of the cortex[18;19]. Our approach adopts a simplified yet effective architecture drawn from this. Rather than separating excitatory and inhibitory neurons, we allow prediction and error correction to occur within a single neuron. By employing two-compartment neurons, our model integrates bottom-up sensory inputs and top-down contextual signals across layers, enabling the dynamic reduction of prediction error in a biologically inspired manner.

To evaluate the effectiveness of the proposed model, we used input stimuli consisting of overlapping digits, creating sensory ambiguity. The network successfully learned to

disambiguate and attend to either the front or back digit depending on an external task cue. Notably, the neuromodulatory mechanism led to measurable changes in firing dynamics, validating its influence on neuronal activity. Moreover, the attentional effects were most pronounced in the deepest hidden layers, aligning with biological evidence that top-down modulation typically emerges in higher-order cortical regions rather than early sensory areas.

2. Materials and methods

2.1. Neuron dynamics and attention model architecture

2.1.1. Network architecture

Our network model adopts a hierarchical architecture in which each hidden layer is connected through both feedforward and feedback pathways (Figure 1A). Each layer represents a distinct cortical region and communicates bidirectionally across layers to exchange bottom-up and top-down input.

The input layer is fully connected to the first hidden layer, delivering forward input in the form of spikes derived from image data. Using rate-based encoding, the image pixel intensities are encoded into spike trains over 50 time steps[20]. This time-varying spike representation mimics biological neuronal communication, where information is conveyed via discrete spikes rather than continuous-valued signals (Figure 1C, 1D.

The first, second, and third hidden layers contain 600, 500, and 500 neurons, respectively. All hidden layers are interconnected through bidirectional feedforward and feedback connections. The output layer comprises 10 non-spiking point neurons, corresponding to the number of target classes. It is fully connected to the highest hidden layer (layer 3), receiving feedforward input from it while sending top-down feedback signals in return. As the final stage for decision-making, the output prediction is determined by the neuron with the highest membrane potential.

In addition to the main network, we introduce attention signal neurons that modulate the excitability of hidden neurons. These signal neurons project to each hidden layer through trainable attention weights and are divided into two groups (5 neurons each), corresponding to the ‘front’ and ‘back’ task cues. They deliver task-specific signals that dynamically adjust the baseline membrane potential of target neurons, thereby enhancing selective responses to task-relevant features in the input.

The attention-modulated spiking activity of each hidden layer is propagated forward to higher layers. Owing to the bidirectional architecture, these modulated signals also shape the top-down feedback transmitted to lower layers. By adjusting neuronal excitability, the model enables task-dependent attention to influence both feedforward and feedback pathways. As the network is trained to minimize energy loss—defined as the discrepancy between the apical and somatic membrane potentials—it learns to optimize synaptic weights while incorporating the effects of attention modulation on neural dynamics.

2.1.2. Neuron dynamics

For the hidden layer neurons, we adopted two-compartment spiking neuron model. They are composed of distinct somatic and dendritic compartments. The dendritic compartment corresponds to the apical tuft, which receives top-down inputs from higher layers, while the somatic compartment integrates bottom-up inputs from lower layers. This neuron model is based on the architecture proposed by Zhang and Bohte[21], who demonstrated that predictive coding behavior can emerge through energy optimization, without requiring explicit error neurons or hard-wired circuits. In their framework, the two-compartment neurons minimize an energy loss defined as the difference between the membrane potentials of the dendritic and somatic compartments, thereby effectively reducing prediction error over time. Building on this simple yet effective architecture, we incorporated a novel attention mechanism into the model. This hierarchical structure allows the model to adopt top-down attention. The following section describes the neuron dynamics used in our model.

The behavior of the hidden neurons follows the Adaptive Leaky-Integrate-and-Fire (ALIF) model, which enhances the standard LIF neuron with adaptive firing mechanisms[22]. The membrane potential dynamics follow:

$$\frac{dV_{a,i}^{l}}{dt} = -\frac{V_{a,i}^{l}}{\tau_a} + \sum_j W_{ij}^{FB} S_j^{l+1}(t) \tag{1}$$

$$\frac{dV_{s,i}^{l}}{dt} = -\frac{V_{s,i}^{l}}{\tau_s} + \sum_j W_{ij}^{FF} S_j^{l-1}(t) + f_{apical}(V_{a,i}^{l}(t)) - b_i^{l}(t) S_i^{l}(t) + \alpha_{att} m_i(t) \tag{2}$$

Here, $V_{a,i}^{l}$ and $V_{s,i}^{l}$ denote the apical and somatic membrane potentials of neuron i in layer l, with τ_a and τ_s as their respective time constants. S_j^{l+1} is the spikes from the layer $l+1$, and W_{ij}^{FB} is the feedback weights from layer $l+1$ to l. Spiking input from the higher layer comes into the apical dendritic part of each hidden neuron.

Spikes from the lower layer S_j^{l-1} is integrated into the somatic voltage with the feedforward weights W_{ij}^{FF} from layer $l-1$ to l. The function f_{apical} determines how apical inputs modulate somatic voltage, defined as:

$$f_{apical}(x) = \frac{1}{2}\left(\frac{1}{1+e^{-x}} - 0.5\right) \tag{3}$$

This function bounds the influence of apical input to the range [-0.25, 0.25]. The adaptive firing threshold makes the neuron more difficult to fire after they activate and is defined as

$$b_i(t) = b_{init} + \beta \eta_i(t) \tag{4}$$

where $b_{init} = 0.1$, $\beta = 1.8$, and $\eta_i(t)$ is an adaptive term governed by:

$$\frac{d\eta_i^l}{dt} = -\frac{\eta_i^l}{\tau_{adp,i}} + S_i^l(t) \tag{5}$$

This term increases upon spiking and decays over time, raising the firing threshold for subsequent spikes. $\tau_{adp,i}$ is the time constant that sets the decay rate of η_i^l , and $S_i^l(t)$ denotes the spikes of neuron i.

$\alpha_{att} m_i(t)$ denotes the attention effect for neuron i and the detailes are explained in the next section.

As noted in Equation 2, the somatic membrane potential decays with the value of $b_i(t)$ which reflects the adaptive effect, after the neuron spikes. A neuron emits a spike $S_i(t) = 1$ when the somatic membrane potential $V_{s,i}^l(t)$ exceeds its firing threshold.

The output neurons evolve according to:

$$\frac{dV_i^{out}}{dt} = -\frac{V_i^{out}}{\tau_{out}} + \sum_j W_{ij}^{FF} S_j^{l-1}(t) \tag{6}$$

Here, V_i^{out} is the membrane potential, and τ_{out} is its decay constant. As the output neurons are non-spiking, feedback to layer 3 is computed using the L2 norm of membrane potentials of the output neurons over time.

2.1.3. Attentional modulation

Inspired by biological attentional modulation processes, we implemented an attention mechanism by dynamically regulating the baseline membrane potential of hidden layer neurons, adopting $\alpha_{att} m_i(t)$ term to somatic voltage 2. This modulation enhances the excitability of neurons associated with relevant features, while suppressing less informative ones, thereby promoting selective and efficient information processing (Figure 1B).

In 2, the modulation effect m evolves over time according to:

$$\frac{dm_i^l}{dt} = -\frac{m_i^l}{\tau_{att}} + \sum_j W_{ij}^{att} S_j^G(t) \tag{7}$$

Here, $m_i(t)$ is the attentional modulation signal for neuron i, driven by spike inputs $S_j^G(t)$ from attention signal neurons. The term τ_{att} (30) controls the temporal decay of the modulation signal which has a larger value than τ_s (15) and τ_a (15). W_{ij}^{att} denotes the trainable synaptic weights projecting from signal neurons to hidden neurons. The influence of $m_i(t)$ is determined by α_{att} (we set $\alpha_{att} = 1.0$).

We set m_i to represent the regulation of ion flow through the ion channels, enhancing the excitability of a neuron when it receives attention and lowering the activity of a neuron when it receives suppression. We designed this attention mechanism to take an effect by regulating baseline membrane potential of each neuron. When a neuron receives a strong positive attention from the signal neurons, $m_i(t)$ increases, which in turn raises the baseline membrane potential, thereby making the neuron easier to fire. Conversely, weak or negative attention signals result in lower baseline membrane potential, making the neuron less likely to fire.

2.2. Training and tasks

2.2.1. Model train

To train the model, we employed surrogate gradient learning using the Multi-Gaussian method[23], which enables gradient-based optimization for spiking neurons. Training was performed using Forward Propagation Through Time (FPTT)[24], allowing online weight updates every K time steps, with both task, energy, E-I ratio, and dynamic regularization losses.

The loss function followed that defined in previous work[21]:

$$\mathcal{L}_t = \alpha_{clf}\mathcal{L}_{clf,t} + \alpha_E\mathcal{L}_{E,t} + \alpha_{EI}\mathcal{L}_{EI,t} + \alpha_{reg}\mathcal{L}_{reg,t} \tag{8}$$

where $\alpha_{clf} = 1.0$, $\alpha_E = 0.05$, $\alpha_{EI} = 0.01$, and $\alpha_{reg} = 1.0$.

Here, $\mathcal{L}_{clf,t}$ is the negative log-likelihood classification loss. $\mathcal{L}_{E,t}$ is the energy loss and defined as:

$$\mathcal{L}_{E,t} = \frac{1}{N}\sum_l\sum_i |V^l_{a,i}(t) - V^l_{s,i}(t)| \tag{9}$$

where N is the total number of hidden neurons. This term captures the average prediction error, as the apical membrane potential V_a represents top-down predictions, while the somatic membrane potential V_s integrates bottom-up sensory evidence. Minimizing this discrepancy enables the network to reduce prediction errors by focusing on task-relevant features in the input.

$\mathcal{L}_{EI,t}$ softly constrains the overall excitatory/inhibitory balance of the network, encouraging approximately 20% of the weights to be negative. We approximated this ratio using a smooth surrogate based on the sigmoid function. Given a weight vector w, we estimated the

proportion of negative weights as:

$$\text{avg\_neg} = \frac{1}{N}\sum_{i=1}^{N}\sigma(-kw_i)$$

where N is the number of weights, σ is the sigmoid function, and k is a steepness parameter (set to 10^3) to sharpen the transition around zero. The penalty was defined as the squared error between this estimated negative ratio and the target value of 0.2:

$$\mathcal{L}_{EI,t} = \text{scale} \cdot (\text{avg\_neg} - 0.2)^2$$

where we used scale $= 5$.

$\mathcal{L}_{reg,t}$ is the dynamic regularization, for stabilizing learning and preventing overfitting. We incorporated a dynamic L2-based regularization term that penalizes deviations from a moving reference value:

$$\mathcal{L}_{reg,t} = \frac{1}{2}\sum_{i}(\theta_i(t) - \theta_{\text{ref},i}(t))^2$$

θ_i(t) is each trainable parameter at time t. $\theta_{\text{ref},i}(t)$ is adaptively updated at each training step using an exponential moving average of the current parameter values, along with a momentum-like term to smooth fluctuations. This dynamic anchoring allows the model to gradually shift away from the initial parameters while still regularizing excessive changes.

The regularization coefficient $\alpha_{\text{reg}}(t)$ decays over time to allow more flexible adaptation in later training stages. We applied a cosine decay schedule to gradually reduce the influence of the regularization term over training epochs. This approach allows the model to benefit from strong regularization in the early stages while gradually relaxing constraints to facilitate task-specific adaptation. The decay function is defined as:

$$\alpha_{\text{reg}}(t) = \alpha_0 \cdot \frac{1}{2}\left(1 + \cos\left(\pi \cdot \frac{t}{T}\right)\right)$$

where t is the current training epoch, T is the total number of epochs, and α_0 is the initial regularization coefficient. This schedule starts at $\alpha_0 = 1.0$ and smoothly decays to zero as training progresses. It helps maintain stability in early epochs while allowing more flexibility in later stages.

2.2.2. Selective attention learning using overlapped digit inputs

To develop a model capable of prioritizing task-relevant components of the input, we designed a classification task using overlapped handwritten digit images. We constructed an

overlapped MNIST dataset by superimposing two translucent digits—one positioned in front and the other behind[25]. We adjusted the transparency levels of the front and back digit to 68% and 38%, respectively. This setup was designed to create the simplest possible scenario that requires selective attention, while ensuring that both digits remain visible without one completely occluding the other.

The model is trained to selectively attend to the task-relevant digit among the two digits presented in one image, based on a classification rule signal delivered by the attention signal neurons. These attention signal neurons provide task-specific cues indicating whether the model should attend to the front or back digit. If the attention mechanism functions as intended and the attention weights are properly learned, the model is expected to adjust its focus accordingly and prioritize the digit corresponding to the instructed classification rule.

Training was conducted in two phases: pretraining and fine-tuning (Figure 2). Both phases used supervised learning with mini-batches (batch size = 200).

In the **pretraining phase**, the model was trained on standard single-digit MNIST images. During this stage, the attention mechanism remained inactive—the attention weights were frozen at their initial values. The model learned basic digit classification by updating only the feedforward and feedback weights.

In the **fine-tuning phase**, the model was trained on the overlapped MNIST dataset, where each image contained two superimposed digits. Here, the network learned to allocate attention appropriately based on the target label (front or back). Attention signal neurons generated Poisson spike trains indicating the task cue, delivering contextual signals to the hidden layers. These signals modulated the baseline membrane potential of hidden neurons, dynamically adjusting their excitability to enhance task-relevant information processing. During this phase, attention weights were trained with a learning rate of 0.001, while the remaining network parameters were updated at a reduced learning rate (0.5×).

Both the pretraining and fine-tuning phases used 60,000 training and 10,000 test images. During fine-tuning, the target digit (front or back) was randomly assigned on a per-sample basis within each batch.

A full list of training parameters is provided in Table 1 and Table 2.

2.3. Model evaluation and analysis

2.3.1. Classification of overlapped MNIST images

To assess whether the model successfully learned to attend to the task-relevant digit and perform accurate classification, we first evaluated its performance on the overlapped MNIST

Parameter	Values
Hidden layer size	600, 500, 500
Number of attention signal neurons	10 (5 front, 5 back)
Input layer size	784
output layer size	10
Epochs	15 (pretrain), 25 (fine-tuning)
Total parameters (pretrain)	1,588,420
Total parameters (fine-tuning)	1,607,620 (all layers att), 1,594,420 (single L att)
Batch size	200
Input time step	50
Drop out	0.4 (pretrain), 0.3 (fine-tuning)
K step	10
Pretrain learning rate	10^{-3}
Fine-tuning learning rate	10^{-3} or 5×10^{-4}
α_{clf}	1
α_E	0.05
α_{reg}	1

Table 1. Hyperparameter for training.

Parameter	Values
Attention sigmoid strength α	3
τ_s	15
τ_a	15
τ_{adp}	20
τ_{att}	30
τ_{out}	5
b_{init}	0.1

Table 2. Hyperparameter for neuron dynamics. Initial values are listed in the table.

dataset. In addition to overall accuracy, we examined incorrect predictions to determine whether the attention mechanism primarily reduced reversed errors—cases in which the model incorrectly classified the distractor digit (e.g., predicting the front digit when the back digit was the target). By comparing how frequently such reversed predictions occurred with and without the attention mechanism, we aimed to determine whether attention effectively guided the model toward the correct target digit.

2.3.2. Mechanisms of attention in the network

We then investigated how the attention mechanism operates within the network. To evaluate whether excitability modulation was successfully implemented, we analyzed the firing activity of hidden layer neurons. Specifically, we examined whether the spiking patterns evoked by overlapped digit images with attention resembled those elicited by standard single-digit images. This comparison allowed us to assess whether attention restored clean representations of task-relevant digits under overlapping conditions.

We also explored how attention interacts with feedback connections by examining changes in feedback signal patterns. To identify which layer receives attention most effectively, we conducted ablation experiments in which attention signals were delivered exclusively to either the first or third hidden layer (L1 or L3), and compared classification performance across conditions.

2.3.3. Preservation of baseline classification ability

Finally, we evaluated whether the trained attention model retained its baseline ability to classify single-digit images. Using the standard MNIST test dataset, we tested the model with fixed synaptic weights and with attention signals turned off. This evaluation ensured that attention training did not impair the model's fundamental digit recognition performance.

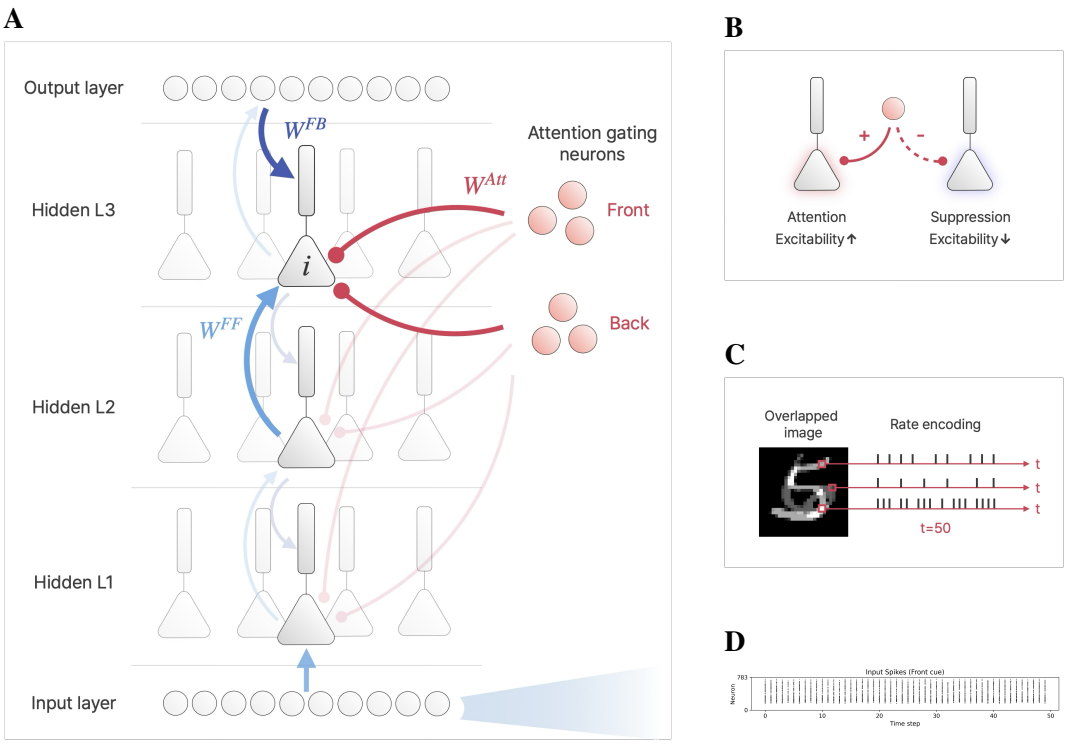

Fig 1. Network architecture. Overall network architecture and input encoding. **(A)** Neurons in the hidden layers, input and output layer are colored light gray, and red colored neurons are attention signal neurons. Dark blue and light blue arrows indicate inter-layer feedback and feedforward connections, respectively. The spikes from the attention signal neurons come into the network with attention weights (red arrows). **(B)** Attention signal neurons regulate excitability of the hidden layer neurons. **(C)** Input images are rate-encoded according to the pixel intensity. **(D)** Sample spike raster plot of an input image of the overlapped MNIST dataset.

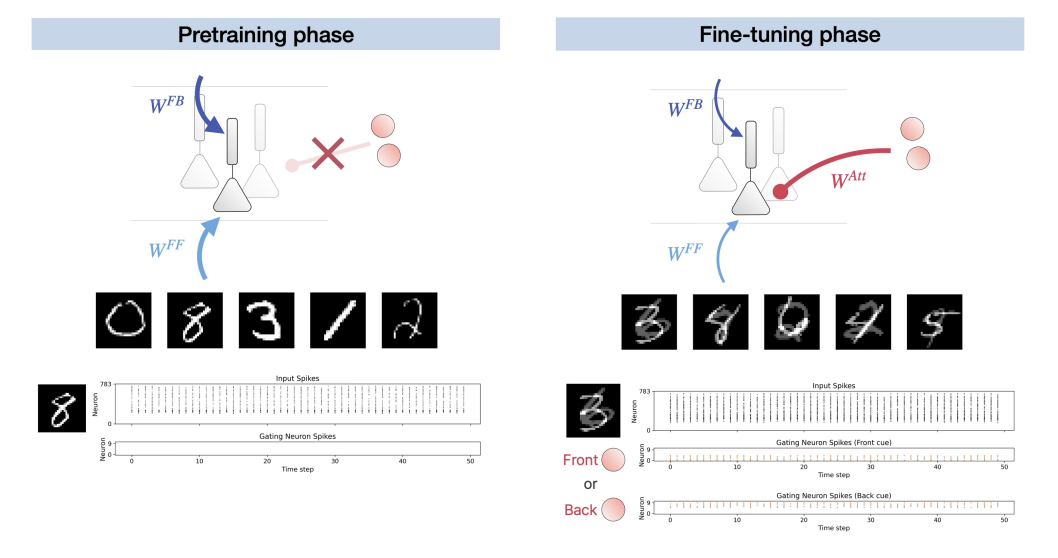

Fig 2. Two training phases. The model undergoes two stages of supervised training to learn selective attention and digit classification. **(Left)** In the pretraining phase, the model is trained on standard MNIST images containing a single digit. During this stage, only the feedforward and feedback weights are updated, while the attention mechanism remains inactive. **(Right)** In the fine-tuning phase, the model is trained on overlapped MNIST images. Attention signal neurons become active and provide task-specific cues (front or back). The model learns to focus on the task-relevant digit by updating attention weights, while other weights continue to be updated with a lower learning rate.

3. Results

3.1. The model could selectively attend to the proper target of ambiguous input

3.1.1. Classification performance and improvement of reversed prediciton in attention model

We first asses the model's classification performance (Figure 3). During the pretraining phase, the model achieved 96.99% classification accuracy after 15 epochs, on standard black-and-white MNIST dataset which contains single digit per image (Figure 3A).

With this pretrained model, we compared the following three models for the fine-tuning phase:

1. Attention model: The attention model presented in this work. Excitability of neurons are modulated, attention signals are presented, and attention weights are properly learned.

2. Random-attention model: Basic frame of the attention model is maintained and attention weights are also learned, but the signal neurons fired randomly, independent of the task cue.

3. Baseline model: No attention mechanism was used; neither signal inputs nor attention were applied.

After fine-tuning phase, the baseline and random-attention attention models reached 57.76%, 57.99% of the best test accuracy respectively, and the attention model achieved 81.67% accuracy (Figure 3B. Compared to the pretrain accuracy, if attention is not applied, the model accuracy drops drastically when overlapped images are presented.

We further examined the reason of incorrect prediction across models. In order to check if attention model has mainly improved its accuracy by choosing the correct digit, we checked how often the model had incorrect prediction is due to focusing on the opposite digit (reversed prediction, e.g., choosing the back digit when the front digit was the target).

The result showed that the main reason for incorrect predictions was due to the reversed prediction for random-attention model and the baseline model. During front-digit classification, the ratio of total incorrect cases was 40.70±0.01% and 33.57±0.01%, and the total reversed prediction ratio was 29.25±0.01% and 28.14±0.01% for the random-attention and baseline model, respectively (Figure 3C). When the back digit was the target, the ratio of total

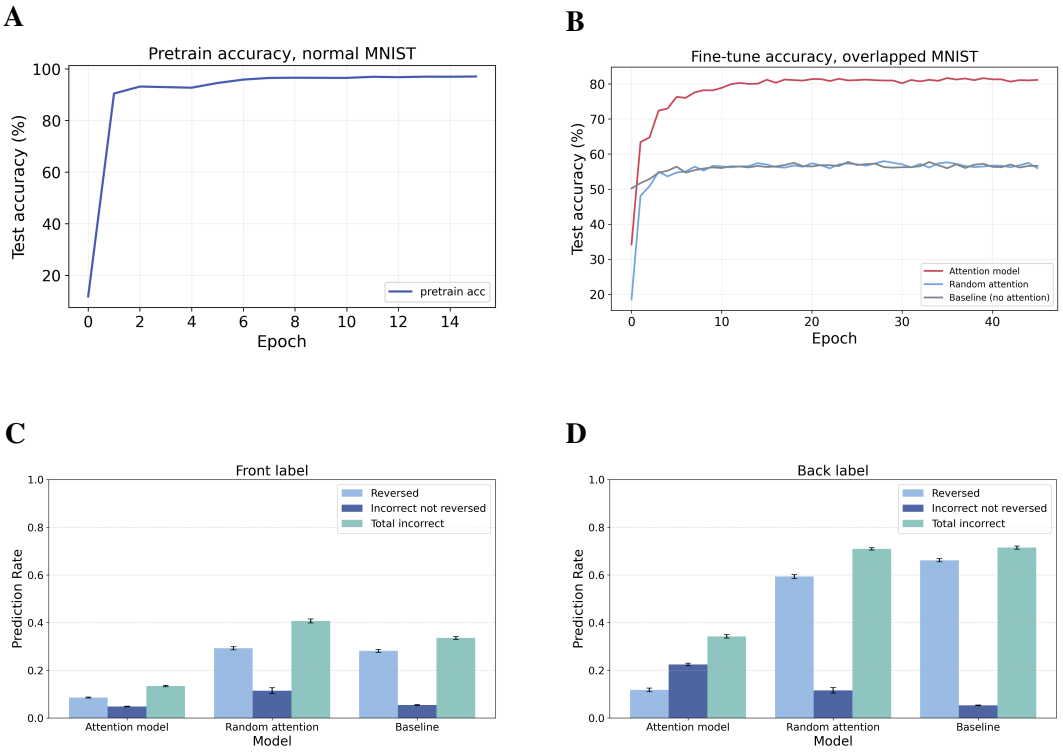

Fig 3. Classification performance for overlapped MNIST dataset. Attention improves recognition of proper target digit for current cue. **(A)** Test accuracy during pretrain. Normal (un-overlapped) black and white MNIST dataset was given as input. In this phase, model learns basic digit patterns. **(B)** Test accuracy during fine-tuning. Attention model is compared with other two settings. Overlapped digit images are given as input. Red line indicates the accuracy of attention model, light blue line indicates accuracy of random-attention model, and gray line indicates accuracy of the baseline model (no attention mechanism implemented). **(C)** Types of incorrect predictions for models, front label. **(D)** Types of incorrect predictions for models, back label **(C, D)** Light blue bar indicates the rate of reversed prediction, dark blue bar indicates the rate of incorrect predictions with not reversed predictions, and the mint color bar is the rate of total incorrect predictions. The reversed prediction (prediction of the opposite digit, e.g. predicting for the back digit when front one is the target) tends to be the main cause of wrong predictions. Attention model significantly improves these errors.

	Attention model	Baseline model
Relaxed Accuracy	88.84 ± 0.22%	91.62 ± 0.49%
True Accuracy (Front digit)	86.94 ± 0.15%	59.88 ± 1.96%
True Accuracy (Back digit)	66.63 ± 0.55%	29.71 ± 1.03%

Table 3. Classification performance in a relaxed condition, regarding correct if the model predicted either digit on front or back.

incorrect cases was 70.71±0.5% and 71.42±0.5%, and the ratio of reversed prediction was 59.19±0.02% and 65.99±0.01% for the random-attention and baseline model, respectively (Figure 3D). Taken together, for front digit classification, the ratio of reversed predictions among the total incorrect predictions was 71.87±0.76%, 83.81±0.96% for random-attention and baseline models. For back digit classification, the ratio of reversed predictions among the total incorrect predictions took upto 83.65±1.64%, 92.55±1.04% for random-attention and baseline models.

In contrast, the attention model showed significant improvements on the reversed prediction (Figure 3C, 3D). When the front digit was the target, the total reversed prediction ratio was 8.56±0.01% and when the back digit was the target, it was 11.57±0.02%.

3.1.2. Basic digit recognition ability in a relaxed condition

To validate the basic classification capability of the baseline model, we also tested them in a relaxed condition where predicting either digit (front or back) was considered correct (Table 3). This helped confirm that the poor performance in the standard setting was primarily due to incorrect attention allocation, not a general failure to recognize digits. Baseline model achieved 91.62 ± 0.49% of relaxed accuracy, and the true accuracy for the front digit (when front digit was the target and the model predicted for the front digit) was 59.88 ± 1.96%. True accuracy for the back digit was 29.71 ± 1.03%. Attention model has a relatively lower relaxed accuracy of 88.84 ± 0.22%. But the model could predict the front digit correctly by 86.94 ± 0.15%, and back digit by 66.63 ± 0.55%.

We next examined if the attention model could preserve its ability to classify single-digit MNIST dataset. The model could predict with the accuracy of 97.02 ± 0.09%, which is similar to the pretrained state, indicating that its basic classification ability is not impaired.

3.2. The model was able to generate internal representation of the correct target input.

3.2.1. Attention model could generate representation of the target digit

We explored how the attention mechanism functions within the network by analyzing neuronal firing in the hidden layer. To determine if excitability modulation was achieved, we looked at whether spiking patterns from each layers matched in two settings—when standard single-digit images are presented and when overlapping digit images are presented along with attention.

We conducted representational similarity analysis (RSA) by computing cosign similarity of the firing patterns of each layer between the two conditions. Figure 4A shows the firing patterns of the attention model has a meaningful similarity between two conditions, successfully generating internal representations of the target digit. The representational similarity for the correct digits tended to get more distinct for deeper layer.

3.2.2. Attention operates with feedback connections

Additionally, we studied the interaction of attention with the feedback connections. When we excluded the feedback signals coming to each hidden layer, the RSA matrices showed less distinct similarity patterns (Figure 4B). In our network architecture, the attention-altered firing patterns are delivered to the superior layers. Therefore, we questioned if it effects the feedback signals, altering the feedback input received in each hidden layer. We assessed changes in feedback signal patterns by calculating L2 norm difference of apical voltage (V_a) patterns in two settings (Figure 4C).

First, we calculated difference between when attention is on and off. In both cases when front digit is the target and the back digit is the target, V_a patterns differed between when attention is on and attention is off. This results show that attention has really changed how hidden layers receive feedback signals. Second, we calculated difference between when attention is on and when attention signals comes along with random noise. In this setting, attention neurons fired indicating the correct sample, but at the same time, the opposite set of the signal neurons also fired with smaller firing probabilities (e.g. when the front digit was the answer, the front signal neurons mainly fired but the back signal neurons also generated weak spike trains). V_a patterns also differed in this setting, showing less difference than when the patterns were compared with or without attention.

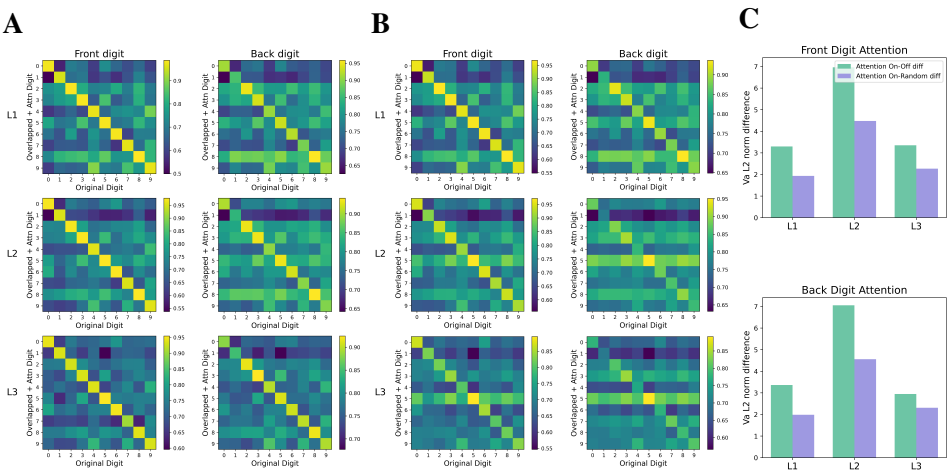

Fig 4. Attention improves representational similarities by altering firing patterns and feedback signals are altered by attention (A, B) Representational similarity matrices, situation between when normal MNIST dataset is given and when overlapped MNIST dataset is given with proper attention signals. All RSA are calculated with attention model which has finished fine-tune training. **(A)** Attention and feedback signals. **(B)** Ablation of feedback signals. **(C)** L2 norm difference of apical voltage patterns. Top: front digit target, Bottom: back digit target. Mint color bars indicate the difference between when attention is on and off. Purple bars indicate the difference between when attention is on and when attention signals comes along with random noise.

3.3. The deepest layer was the most effective target of attention

3.3.1. Attention Weights Focus on Layer 3

We first examined how attention weights evolved during the fine-tuning phase. Figure 5 attention weight matrices of each layer across training epochs. As training progressed, the connection patterns from the front and back signal neurons began to diverge, suggesting that the model was learning to differentiate attention signals based on task demands. This divergence was most pronounced in layer 3, the deepest hidden layer, indicating that it played a central role in modulating neural activity according to the task cues. The distinct separation in weight patterns suggests that layer 3 played a key role in task-specific modulation and served as providing top-down attention signals to the shallower layers.

3.3.2. Layer-wise Impact of Attention on Neural Activity

We further investigated whether the impact of attention on neural firing patterns also varied across layers. To do this, we compared hidden layer activity in the attention model under two conditions: when the correct attention signal was provided versus when an incorrect (reversed) attention signal was given. Figure 6A (right) presents the L2 norm of the difference in firing patterns between these two conditions. For both front and back digit tasks, the discrepancy increased progressively from layer 1 to layer 3, suggesting that deeper layers were more strongly modulated by attention. We also examined the effect of noisy attention signals—when the attention cue was randomly corrupted. As shown in Figure 6A (left), although the difference between the correct and noisy conditions was smaller than in the reversed condition, the same trend persisted: layer 3 consistently exhibited the largest changes in firing patterns, indicating a greater sensitivity to attention signals.

In Figure 6B, we quantified the layer-wise difference in firing rates as a signed difference between the attention-on and attention-off conditions. In both the false and noisy attention cases, the deviation in firing rate became larger in deeper layers, again supporting the idea that attention exerts a progressively stronger influence toward layer 3.

3.3.3. Layer 3 Attention Yields Most Efficient Learning

To determine which layer benefits most from attention, we conducted an ablation study in which attention was applied exclusively to either layer 1 or layer 3. Figure 6C compares the energy loss during fine-tuning for three models: one with attention applied to all hidden layers, one with attention only to layer 3 (L3-attention model), and one with attention only

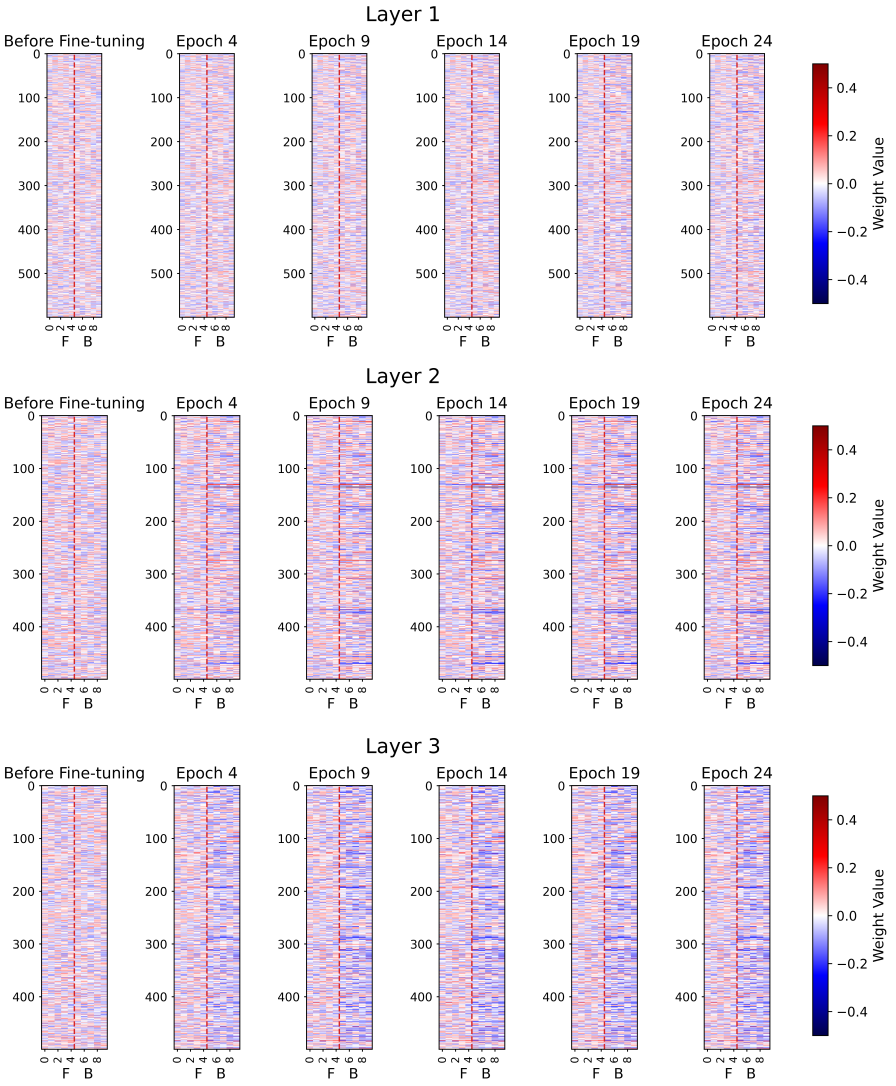

Fig 5. Attention weight matrices for three hidden layers (log-scale). Attention weights evolve differently across layers during training, diverging pattern most profound in the deepest layer.

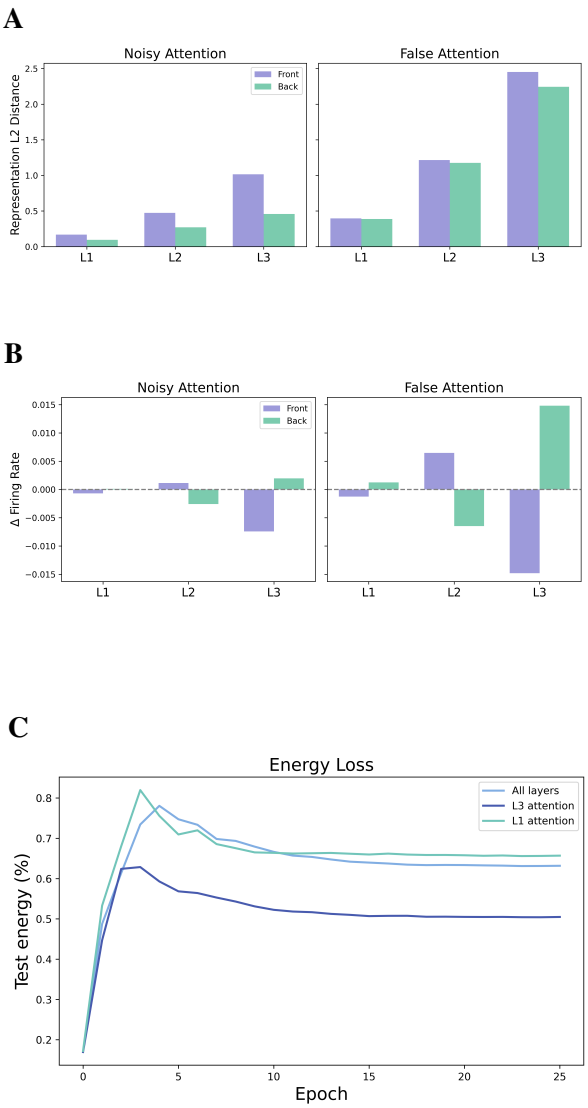

Fig 6. Layer-wise analysis of attention and performance. **(A)** L2 norm Differences in firing activity patterns between attention conditions. **(B)** Signed difference of firing patterns. **(C)** Energy loss during training, for three models with ablation of layer 1 or layer 3.

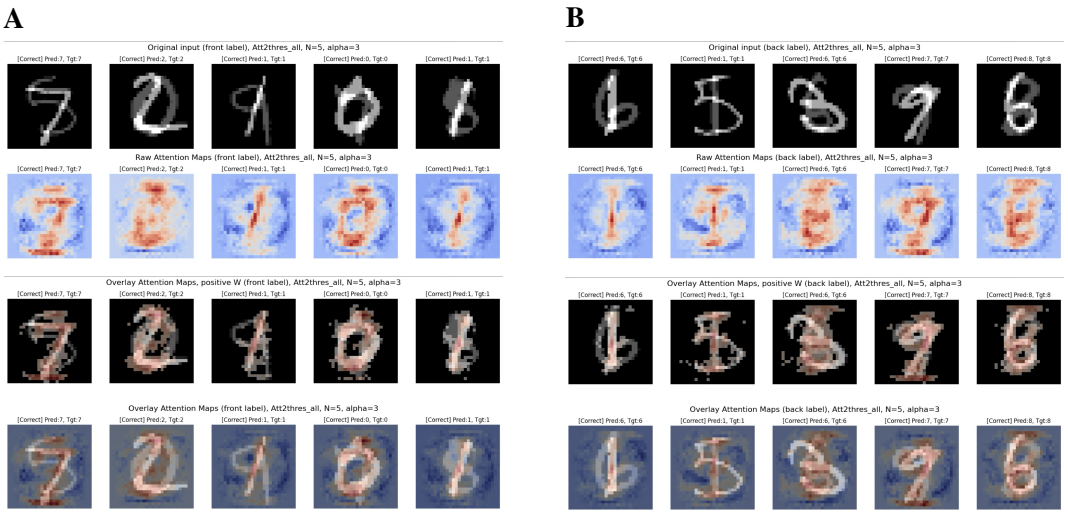

Fig 7. Attention map. Attention via threshold modulation enables the model to selectively focus on the task-relevant digit. The first row displays the input images. The second row shows the raw attention maps. The third row overlays thresholded attention maps (> 0.3) onto the inputs. The fourth row overlays the same maps with alpha=0.5 for visualization. **(A)** Task requires attending to the front digit. **(B)** Task requires attending to the back digit.

to layer 1 (L1-attention model). The L3-attention model achieved the lowest final energy loss, suggesting that directing attention to layer 3 yields the most efficient learning outcome.

3.4. Spatial Localization of Attention

To understand how attention was spatially distributed during fine-tuning, we visualized the attention maps generated by the model. Figure 7 shows both the standalone attention maps and the same maps overlaid on the corresponding input images.

The attention model effectively allocated its focus to the digit region relevant to the current task label—either the front or back digit—demonstrating that the attentional modulation was both spatially selective and task-dependent. Positive attention was concentrated on the target digit while suppressing focus on irrelevant background areas.

One limitation we observed was a tendency for the model to focus disproportionately on the front digit when it was a ‘1’, even in tasks where the back digit was the correct target. Nevertheless, even in such cases, the model still assigned weak but meaningful attention to

the correct digit and was able to make accurate predictions.

The attention maps were computed using the spike-firing-rate (SFR) method. See Appendicies 3 for details.

4. Discussion

In this study, we investigated how neuromodulatory principles—particularly inspired by cholinergic attentional modulation—can be effectively integrated into a spiking neural network (SNN) using dynamic threshold control. By implementing task-dependent attention gating neurons, we enabled the model to adaptively regulate neuronal excitability across hidden layers, leading to enhanced classification of ambiguous, overlapped visual stimuli.

4.1. Attention improves task-relevant representation under ambiguity

Our results show that the attention model significantly outperformed the baseline and random-attention models in classifying overlapped digit images. Notably, the main source of errors in non-attention models was the reversed prediction—choosing the distractor digit instead of the task-relevant one. This pattern was drastically reduced in the attention model, suggesting that the implemented attention mechanism effectively directed the model's focus to the appropriate digit, in line with the intended top-down cue. These findings support the view that attentional modulation enables more accurate resolution of perceptual ambiguity by biasing internal processing toward task-relevant features.

4.2. Attention shapes internal representations and feedback dynamics

We further observed that attentional modulation induced meaningful changes in neuronal firing patterns across layers. Representational similarity analysis (RSA) demonstrated that the attention model's responses to overlapped inputs resembled those for clean, single-digit inputs, particularly in deeper layers. This implies that attention helped the network reconstruct or emphasize the target digit's internal representation despite the presence of distractors.

Additionally, we found that attention altered feedback signals transmitted via the apical dendrites. Apical membrane potentials changed significantly depending on the attention condition (on vs. off, or true vs. noisy signal), indicating that top-down feedback itself was modulated by attention. This dynamic interaction between attention and feedback supports the predictive coding framework, wherein top-down signals are shaped to reduce prediction errors and refine perception.

4.3. Layer 3 is the most effective site for attentional modulation

A layer-wise analysis revealed that attention exerted the strongest influence in the deepest hidden layer (layer 3). Attention weight matrices showed the most prominent separation between front and back gating signals in layer 3, and firing rate modulation due to attention was also greatest in this layer. Furthermore, when attention was restricted to a single layer, models with attention applied only to layer 3 achieved the lowest energy loss and better performance than those with attention applied to earlier layers. This suggests that deeper layers are more receptive to attentional signals and more capable of integrating them into meaningful modulation of network activity.

This finding resonates with biological evidence suggesting that higher cortical areas (e.g., prefrontal cortex) play a central role in attention control by sending task-specific modulatory signals to earlier sensory areas. In our model, layer 3 may serve a similar role by acting as an internal "attention hub" that integrates task cues and distributes attentional influence downstream through both feedforward and feedback pathways.

4.4. Limitations and future directions

While our attention model successfully improved classification under ambiguous conditions, certain limitations remain. The model occasionally exhibited a bias toward more salient digits, such as the digit '1' when presented in the front, regardless of the task cue. This suggests a possible imbalance in the saliency-driven bottom-up input that may compete with top-down modulation. Addressing such biases may require integrating additional mechanisms such as normalization, inhibitory control, or uncertainty-based attention scaling.

Moreover, although our attention mechanism mimics biological principles through dynamic threshold modulation, it does not yet incorporate more complex neuromodulatory dynamics, such as context-sensitive gain control or interactions between multiple neurotransmitter systems. Future work could explore more biologically grounded attention circuits, or implement learning rules that adapt attention weights in a task-specific and energy-efficient manner.

Finally, extending this model to naturalistic or sequential data, and evaluating generalizability beyond synthetic overlapping digits, would help assess the broader applicability of the attention mechanism.

5. Conclusion

This study proposed a biologically inspired attention mechanism for spiking neural networks (SNNs), implemented via dynamic modulation of neuronal firing thresholds. By drawing on neuromodulatory principles—particularly those associated with cholinergic attentional control—we designed a task-driven attention gating system that selectively regulated the excitability of hidden neurons.

Through experiments using overlapped digit classification, we demonstrated that the attention model substantially outperformed baseline and random-attention models, particularly by reducing errors caused by misdirected attention. The model exhibited improved focus on the task-relevant digit and showed enhanced internal representations that closely resembled clean inputs. Layer-wise analyses revealed that the deepest hidden layer (layer 3) served as the most effective site for attention integration, showing the greatest modulation in firing patterns and energy optimization.

These findings highlight the computational benefits of integrating biologically grounded attention mechanisms into hierarchical SNN architectures. Our work bridges predictive coding principles with neuromodulatory attention control, and provides a promising framework for building more flexible and context-sensitive spiking models. Future directions include extending this model to naturalistic inputs, exploring multi-modal attention systems, and developing learning rules that dynamically adapt attentional gain based on task uncertainty and context.

References

1. Desimone R, Duncan J, et al. Neural mechanisms of selective visual attention. Annual review of neuroscience. 1995;18(1):193–222.

2. Briggs F, Mangun GR, Usrey WM. Attention enhances synaptic efficacy and the signal-to-noise ratio in neural circuits. Nature. 2013;499(7459):476–480.

3. Knudsen EI. Neural circuits that mediate selective attention: a comparative perspective. Trends in neurosciences. 2018;41(11):789–805.

4. Thiele A, Bellgrove MA. Neuromodulation of attention. Neuron. 2018;97(4):769–785.

5. Ferguson KA, Cardin JA. Mechanisms underlying gain modulation in the cortex. Nature Reviews Neuroscience. 2020;21(2):80–92.

6. Noudoost B, Moore T. The role of neuromodulators in selective attention. Trends in cognitive sciences. 2011;15(12):585–591.

7. Kanamaru T, Aihara K. Acetylcholine-mediated top-down attention improves the response to bottom-up inputs by deformation of the attractor landscape. PloS one. 2019;14(10):e0223592.

8. McCormick DA, Prince DA. Mechanisms of action of acetylcholine in the guinea-pig cerebral cortex in vitro. The Journal of physiology. 1986;375(1):169–194.

9. Nadim F, Bucher D. Neuromodulation of neurons and synapses. Current opinion in neurobiology. 2014;29:48–56.

10. Marder E. Neuromodulation of neuronal circuits: back to the future. Neuron. 2012;76(1):1–11.

11. Eggermann E, Feldmeyer D. Cholinergic filtering in the recurrent excitatory microcircuit of cortical layer 4. Proceedings of the National Academy of Sciences. 2009;106(28):11753–11758.

12. Xiao Y, Yuan Q, Jiang K, Zhang Q, Zheng T, Lin CW, et al. Spiking Meets Attention: Efficient Remote Sensing Image Super-Resolution with Attention Spiking Neural Networks. arXiv preprint arXiv:250304223. 2025;.

13. Zhou Z, Che K, Fang W, Tian K, Zhu Y, Yan S, et al. Spikformer v2: Join the high accuracy club on imagenet with an snn ticket. arXiv preprint arXiv:240102020. 2024;.

14. Yao M, Zhao G, Zhang H, Hu Y, Deng L, Tian Y, et al. Attention spiking neural networks. IEEE transactions on pattern analysis and machine intelligence. 2023;45(8):9393–9410.

15. Yao M, Hu J, Zhou Z, Yuan L, Tian Y, Xu B, et al. Spike-driven transformer. Advances in neural information processing systems. 2023;36:64043–64058.

16. Larkum ME, Senn W, Lüscher HR. Top-down dendritic input increases the gain of layer 5 pyramidal neurons. Cerebral cortex. 2004;14(10):1059–1070.

17. Sripati AP, Johnson KO. Dynamic gain changes during attentional modulation. Neural Computation. 2006;18(8):1847–1867.

18. Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ. Canonical microcircuits for predictive coding. Neuron. 2012;76(4):695–711.

19. Hertäg L, Sprekeler H. Learning prediction error neurons in a canonical interneuron circuit. Elife. 2020;9:e57541.

20. Eshraghian JK, Ward M, Neftci E, Wang X, Lenz G, Dwivedi G, et al. Training spiking neural networks using lessons from deep learning. Proceedings of the IEEE. 2023;111(9):1016–1054.

21. Zhang M, Chitic R, Bohté SM. Energy optimization induces predictive-coding properties in a multi-compartment spiking neural network model. PLOS Computational Biology. 2025;21(6):e1013112.

22. Bellec G, Scherr F, Subramoney A, Hajek E, Salaj D, Legenstein R, et al. A solution to the learning dilemma for recurrent networks of spiking neurons. Nature communications. 2020;11(1):3625.

23. Yin B, Corradi F, Bohté SM. Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks. Nature Machine Intelligence. 2021;3(10):905–913.

24. Kag A, Saligrama V. Training recurrent neural networks via forward propagation through time. In: International Conference on Machine Learning. PMLR; 2021. p. 5189–5200.

25. O'Craven KM, Downing PE, Kanwisher N. fMRI evidence for objects as the units of attentional selection. Nature. 1999;401(6753):584–587.

Abstract in Korean

표준 미세 회로 기반 주의 스파이킹 신경망 모델

선택적 주의(selective attention)는 모호한 환경 속에서 목표와 관련된 자극을 우선적으로 처리할 수 있도록 하는 핵심적인 인지 기능이다. 본 연구는 콜린성 주의 조절을 포함한 신경조절 메커니즘(neuromodulation)에서 영감을 받아, 뉴런의 흥분도를 동적으로 조절하는 생물학적으로 타당한 주의 메커니즘을 스파이킹 신경망(Spiking Neural Network, SNN)에 구현하였다. 제안된 모델은 두 구획(two-compartment) 구조의 뉴런으로 구성된 세 개의 은닉층을 포함하며, 과제 단서(cue)에 따라 입력 이미지의 전경 혹은 배경 숫자에 주의를 집중하도록 설계되었다. 학습은 일반 숫자 이미지로 기본적인 분류 능력을 학습하는 사전 학습 단계와, 겹쳐진 숫자 이미지와 주의 단서를 함께 제공하여 주의 조절을 학습하는 미세 조정 단계로 구성되었다. 실험 결과, 주의 메커니즘을 적용한 모델은 주의를 사용하지 않은 기본 모델 및 무작위 주의 모델에 비해 높은 정확도를 보였으며, 과제와 무관한 숫자에 반응하는 '반전 오류'를 현저히 줄였다. 특히 가장 깊은 은닉층(3 층)에서 주의 조절의 효과가 가장 뚜렷하게 나타났으며, 에너지 사용 효율도 높아졌다. 또한 억제를 억제하는(disinhibition) 방식의 주의 조절이 해당 층에서 주로 이루어졌으며, 주의가 과제와 관련된 숫자 영역에 정확하게 집중된다는 점도 확인되었다. 본 연구는 생물학적 신경조절 원리를 바탕으로 스파이킹 신경망에서 과제 기반 선택적 처리를 효과적으로 구현할 수 있음을 보여주며, 향후 복잡한 환경이나 다양한 주의 과제를 다루는 신경망 설계에 기초적인 방향을 제시한다.

핵심되는 말: 예측 코딩 이론, 주의 회로, 스파이킹 신경망, 분류 학습