Long delays reduce the need for precise weights in spiking neural networks

Pengfei Sun[1], Jascha Achterberg[2], Dan F. M. Goodman[1], Danyal Akarca[1,3*]

[1]Department of Electrical and Electronic Engineering, Imperial College London
[2]Centre for Neural Circuits and Behaviour, University of Oxford
[3]MRC Cognition and Brain Sciences Unit, University of Cambridge

Abstract

Recent work has shown that the performance of spiking neural networks (SNNs) on temporally complex tasks improves significantly when axonal delays are treated as learnable parameters. This raises an important question: If temporal delays improve a network's computational capacity, how precise do synaptic weights need to be? In this work, we investigate the relationship between delay-based computation and weight precision by combining quantized synaptic weights with a range of learnable delays on a challenging neuromorphic audio task. Our results reveal that short delays contribute little to performance, whereas medium to long delays are critical. Building on this insight, we introduce a learnable thresholding mechanism to suppress short delays that can be effectively compensated for by weights. These findings suggest that delays can reduce the burden on weight precision, highlighting a promising direction for energy-efficient SNN design and offering new perspectives on the role of delay in biological and neuromorphic computation.

Keywords: Axonal delays, Delay-based computation, Spiking neural networks, Neuromorphic computing, Supervised learning

Introduction

Spiking neural networks (SNNs) have garnered increasing attention due to their inherent sparsity, energy efficiency, and biological plausibility. Recent work has shown that incorporating trainable axonal delays can significantly boost performance on tasks with rich temporal structure (Sun, Zhu, & Botteldooren, 2022; Hammouamri, Khalfaoui-Hassani, & Masquelier, n.d.; D'agostino et al., 2024). In biological systems, such delays—reflecting neural heterogeneity—are known to expand memory capacity and improve robustness in dynamic environments (Perez-Nieves, Leung, Dragotti, & Goodman, 2021).

While delay learning holds promise as both a primary and auxiliary computational mechanism, its exact role in shaping network performance remains underexplored. In particular, it is unclear whether short or long delays are more critical for solving complex tasks.

In this paper, we present the first study to combine quantized synaptic weights with trainable delays, leveraging this integration to tackle the challenging Spiking Heidelberg Digits (SHD) auditory benchmark (Cramer, Stradmann, Schemmel, & Zenke, 2020). We first replicate the finding that introducing trainable delays substantially improves performance, achieving highly competitive results to the state-of-the-art. To better understand the computational role of delay, we systematically ablate delays from the final layer of the network. Our findings reveal that short delays provide little contribution, while medium to long delays are essential to maintain high accuracy. Based on these insights, we propose a learnable thresholding mechanism that selectively prunes short delays, enabling their effects to be absorbed by the synaptic weights.

Methods

In this study, we use a feedforward spiking neural network (SNN) with trainable axonal delays, following the approach introduced in (Sun et al., 2022), and optimize it end-to-end using the Slayer framework (Shrestha & Orchard, 2018) with finite gradient approximations. The network architecture consists of two fully connected layers with 128 neurons each, and is evaluated on the Spiking Heidelberg Digits (SHD) classification task. The SHD dataset converts spoken digits in English and German into spike trains using a biologically inspired cochlear model, resulting in input across 700 frequency channels. The task involves classifying 20 words based on this spatiotemporal input.

We employ the spikemax loss function (Shrestha, Zhu, & Sun, 2022), and final predictions are made by selecting the output neuron with the highest cumulative spike count. The spiking units are modeled using the Spike Response Model (SRM), and spike generation is approximated using surrogate gradient descent. Importantly, no upper bound is imposed on delay values during training, allowing the network to freely explore temporal strategies.

To investigate the interaction between delays and weight precision, we quantize synaptic weights into ternary states (Li, Liu, Wang, Zhang, & Yan, 2016) – explained below – and train them using the straight-through estimator. We further introduce a learnable thresholding mechanism to filter out short delays. For each layer, an independent parameter θ_d defines the minimum effective delay, such that any delay d below this threshold is suppressed: $d \leftarrow d \cdot H(d - \theta_d)$, where H is the Heaviside function. This mechanism enables the network to dynamically ignore delays that can be compensated for by the neural weight.

Results

We first aimed to investigate to what extent quantization of the weights impacted task performance in the SHD task. In what follows, "full-precision weights" refers to weights represented using 32 bits, while "quantized weights" refers to $1.58 = \log_2 3$ bits, where the weights are represented only

*Corresponding author: d.akarca@imperial.ac.uk

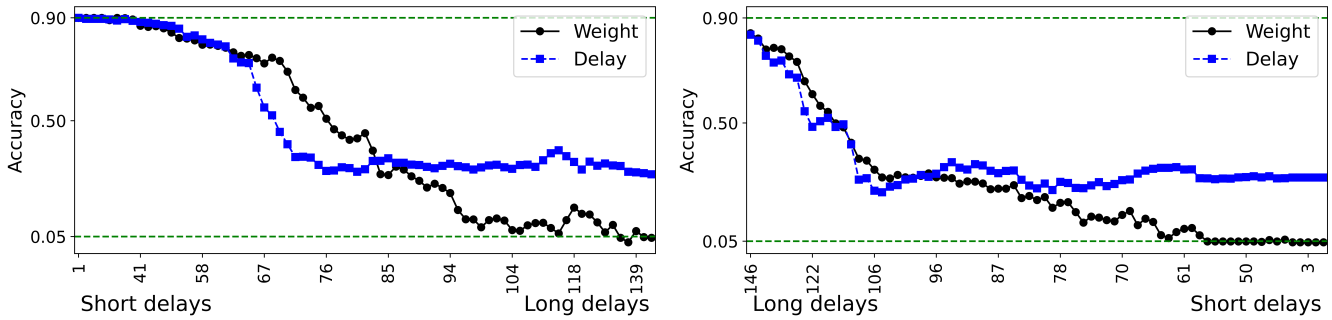

Figure 1: Network accuracy, ranging from baseline performance to chance level, is shown as we ablate k units ranked by increasing delay values (panel left), as well as under the reverse order condition (panel right). The results are presented for two resetting schemes: in the "weight" scheme, the connection weights associated with a specified delay (x-axis value) are set to zero, whereas in the "delay" scheme, the neuron's delay is reset to zero. The removal of long delays (right) has a much larger impact on performance than the removal of short delays (left) in both schemes.

Table 1: Comparison of model performance on the SHD dataset.

Model	**Accuracy**
Full Precision SNN	48.60%
Quantized SNN	44.41%
Quantized SNN+delays	89.92%
Quantized SNN+delays (threshold)	90.68%

as a single negative value, positive value, or zero (otherwise known as ternary).

In the absence of delay learning, full-precision feedforward SNN and its quantized counterpart achieve relatively low accuracies of 48.60% and 44.41%, respectively. For networks operating on temporally complex data, simply increasing the number of neurons for quantization SNNs does not yield substantial performance gains; when the neuron count is increased to 256 and 512, the accuracies obtained are only 45.89% and 45.23%, respectively. However, we find the striking result that, in the presence of trainable axonal delays, it is possible to achieve a significantly higher accuracy of 89.92% even with weights quantized down to ternary (see Table 1). To the best of our knowledge, only one other study has investigated quantization in delay-based networks, with fixed/ternary weights and on a small-scale image classification task (Grappolini & Subramoney, 2023).

This result suggests that delays can substantially enhance the computational capacity of spiking networks. However, it remains unclear precisely *where* in the delay distribution this improvement originates. To investigate this, we fixed the most high-performing of our spiking networks and systematically ablated subsets of delays using two methods: (1) zeroing the connection weights associated with specific delays, and (2) zeroing the delays of the neurons corresponding to those delays. As shown in Figure 1, removing short delays had minimal impact on performance, with competitive accuracy maintained across both ablation strategies. In contrast, removing long delays led to a sharp degradation in accuracy. Intuitively, when all delays were set to zero, the model collapsed to a standard feedforward SNN with significantly reduced performance.

These results demonstrate that even with quantized weights, delay-based SNNs were capable of solving temporally demanding tasks. Motivated by the limited utility of short delays, we introduced a learnable delay threshold to filter them out during training. This dynamic filtering allowed the model to retain temporal expressivity while reducing reliance on fine-grained delay tuning, which can be regarded as a regularizer. The loss in temporal precision was effectively compensated by synaptic weight adjustments, leading to a slight improvement in overall performance. The resulting delay distribution is visualized in Figure 2.

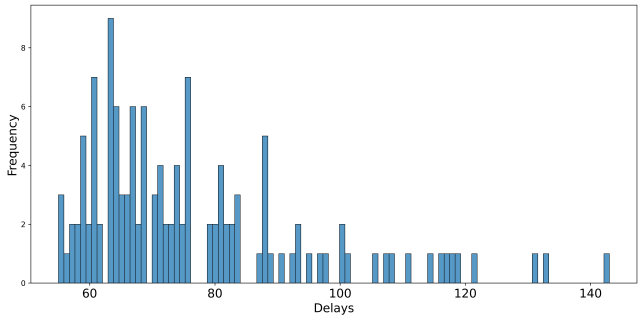

Figure 2: Following training, we observe a positively-skewed delay distribution under learnable minimum threshold constraints.

Conclusions

Delays are a deeply embedded property of the brain's organization (Sreenivasan & D'Esposito, 2019). Here, we show that incorporating trainable axonal delays into a feedforward SNNs can significantly enhances performance on the challenging SHD classification task, achieving nearly 90% accuracy—compared to below 50% for baseline models without delay learning, despite their weights being significantly quantized. Ablation studies revealed that short delays contribute minimally, whereas medium and long delays are essential for high performance. Additionally, introducing a learnable delay threshold to suppress ineffective short delays led to a modest performance gain. In future work, we plan to explore additional constraints, such as sparsity and space (Achterberg, Akarca, Strouse, Duncan, & Astle, 2023), to further refine the delay distribution — potentially yielding models that are both more biologically plausible and hardware-efficient.

References

Achterberg, J., Akarca, D., Strouse, D., Duncan, J., & Astle, D. E. (2023). Spatially embedded recurrent neural networks reveal widespread links between structural and functional neuroscience findings. *Nature Machine Intelligence*, *5*(12), 1369–1381.

Cramer, B., Stradmann, Y., Schemmel, J., & Zenke, F. (2020). The heidelberg spiking data sets for the systematic evaluation of spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(7), 2744–2757.

D'agostino, S., Moro, F., Torchet, T., Demirağ, Y., Grenouillet, L., Castellani, N., ... Payvand, M. (2024). Denram: neuromorphic dendritic architecture with rram for efficient temporal processing with delays. *Nature communications*, *15*(1), 3446.

Grappolini, E., & Subramoney, A. (2023). Beyond weights: deep learning in spiking neural networks with pure synaptic-delay training. In *Proceedings of the 2023 international conference on neuromorphic systems* (pp. 1–4).

Hammouamri, I., Khalfaoui-Hassani, I., & Masquelier, T. (n.d.). Learning delays in spiking neural networks using dilated convolutions with learnable spacings. In *The twelfth international conference on learning representations.*

Li, F., Liu, B., Wang, X., Zhang, B., & Yan, J. (2016). Ternary weight networks. *arXiv preprint arXiv:1605.04711*.

Perez-Nieves, N., Leung, V. C., Dragotti, P. L., & Goodman, D. F. (2021). Neural heterogeneity promotes robust learning. *Nature communications*, *12*(1), 5791.

Shrestha, S. B., & Orchard, G. (2018). Slayer: Spike layer error reassignment in time. *Advances in neural information processing systems*, *31*.

Shrestha, S. B., Zhu, L., & Sun, P. (2022). Spikemax: spike-based loss methods for classification. In *2022 international joint conference on neural networks (ijcnn)* (pp. 1–7).

Sreenivasan, K. K., & D'Esposito, M. (2019). The what, where and how of delay activity. *Nature reviews neuroscience*, *20*(8), 466–481.

Sun, P., Zhu, L., & Botteldooren, D. (2022). Axonal delay as a short-term memory for feed forward deep spiking neural networks. In *Icassp 2022-2022 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 8932–8936).