Generalizable, real-time neural decoding with hybrid state-space models

Avery Hee-Woon Ryoo* **(hee-woon.ryoo@mila.quebec)**
Mila – Quebec AI Institute
Université de Montréal

Nanda H Krishna* **(nanda.harishankar-krishna@mila.quebec)**
Mila – Quebec AI Institute
Université de Montréal

Ximeng Mao* **(ximeng.mao@mila.quebec)**
Mila – Quebec AI Institute
Université de Montréal

Matthew G Perich† **(matthew.perich@umontreal.ca)**
Mila – Quebec AI Institute
Université de Montréal

Guillaume Lajoie† **(g.lajoie@umontreal.ca)**
Mila – Quebec AI Institute
Université de Montréal
Canada CIFAR AI Chair

*equal contribution †equal advising

Abstract

Brain-computer interfaces (BCIs) offer a promising approach for restoring mobility and communication in individuals with paralysis, motor impairments, or neurodegenerative diseases. This is achieved by learning a "decoder", which translates neural activity to an intended behaviour. Traditional decoding approaches often rely on simple statistical methods and recurrent neural networks that are highly specific to individual sessions of data collection, resulting in models that struggle to generalize to new data. On the other hand, empowered by the availability of large-scale multi-session neural datasets, recently developed Transformer-based decoders demonstrate strong generalization performance, but are ill-suited for real-time inference due to their high computational complexity. To bridge the gap between these two approaches, we propose POSSM, a hybrid architecture that combines attention with a state-space model backbone. Our trained models demonstrate efficient real-time inference on intervals that are a fraction of a second long, while also demonstrating strong generalization to unseen sessions and individuals, thereby preserving the strengths of both the traditional and modern approaches to neural decoding.

Keywords: neural decoding; brain-computer interfaces; neural spiking data; state-space models; attention

Introduction

Neural decoding—the process of mapping neural activity to behavioural or cognitive variables—remains a core objective for brain-computer interfaces (BCIs). Through advances in electrophysiological recording techniques and artificial intelligence (AI), recent years have seen great strides in the design of neuroprosthetics for decoding movement (Flesher et al., 2021) and speech (Willett et al., 2023), as well as understanding memory retrieval (Chandravadia et al., 2020).

In particular, motor BCIs can provide a more naturalistic approach to restoring movement by predicting cursor movements intended by the individual, given concurrently recorded neural activity. Along with the development of neuroprosthetics, it can provide a solution for patients suffered from paralysis or amputation (Collinger et al., 2013).

In this paper, we focus on neural spiking data in motor tasks, one of the most commonly encountered modalities for brain datasets. Spikes are irregularly-timed electrical pulses neurons fire to communicate with one another. From neuroscience experiments, spiking events associated with individual units can be extracted from recorded neural activity and used for neural decoding. Traditionally, neural decoding methods rely on "binning", a process where the spikes occurring within each time bin are counted for individual units and used as input to the decoder. Typical decoder models include statistical methods (Wu et al., 2002) as well as recurrent and feedforward neural networks (Glaser et al., 2020). However, many of these methods hard-code the number of units in their input dimension, making it cumbersome both when aggregating multiple sessions with different combinations of units and when transferring to new sessions. These shortcomings have motivated an approach that is more expressive and generalizable.

Building upon recent advancements in deep learning (Jaegle et al., 2022), Azabou et al. (2023) have recently proposed a scalable Transformer-based framework for neural decoding, called POYO. In contrast to binning-based approaches, POYO is equipped with a new tokenization scheme that treats individual spikes as tokens. As a result, POYO does not suffer from loss of temporal resolution as with binning, nor does it rigidly restrict the input format in terms of the number of units. Pre-trained on large-scale multi-session neural datasets, it demonstrated state-of-the-art generalization performance on held-out sessions and even new animals and datasets. However, POYO is non-causal, therefore requiring full 1-second context windows of spike trains, which in turn impose a latency of the same duration for behaviour decoding. In addition, the use of stacked self-attention layers in the model leads to an expensive inference complexity that scales quadratically with respect to the number of latent tokens extracted from each context window. While this number is held constant in POYO for a context length of 1 second, more latent tokens are likely required for longer time sequences. These attributes largely limit its real-time decoding applications.

To overcome these challenges, we propose a hybrid architecture called POSSM, which retains both the flexibility of the POYO tokenizer and a much faster inference speed brought by the state-space model (SSM) backbone. We demonstrate that POSSM can be trained on large-scale multi-session neural datasets spanning multiple animals, tasks and datasets, while also generalizing to new sessions—all with considerably smaller 50ms windows. Moreover, POSSM benefits from a computational complexity comparable to that of recurrent networks, marking substantial advantages over the POYO model in practice. These characteristics make POSSM an ideal candidate for real-time, online neural decoding.

Methods

POSSM is a hybrid model comprising attention and recurrence, and is amenable to real-time neural decoding. As shown in Fig. 1, POSSM consists of a tokenizer for neural spikes, a cross-attention module to construct latent representations of this spiking activity, and an SSM backbone to process sequences of such latents and decode behaviour.

Spike Tokenization. We adopt the tokenization scheme from POYO, shown in Fig. 2. Each spike acts as a token containing information on the neural unit it came from and the timestamp at which it occurred. This is achieved by representing each neural unit as a learnable "unit embedding" and augmenting it element-wise with a rotary positional embedding (RoPE) (Su et al., 2023) associated with the timestamp.

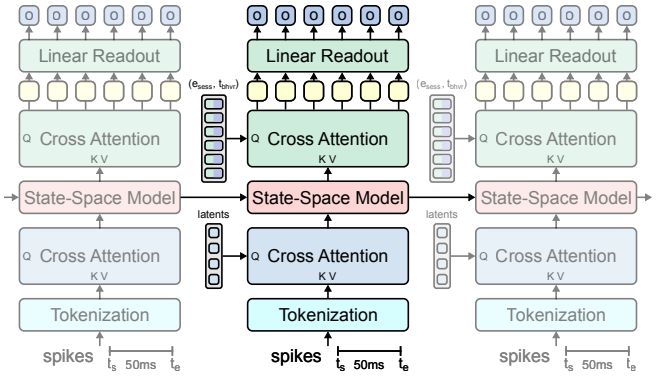

Figure 1: Proposed POSSM architecture for neural decoding.

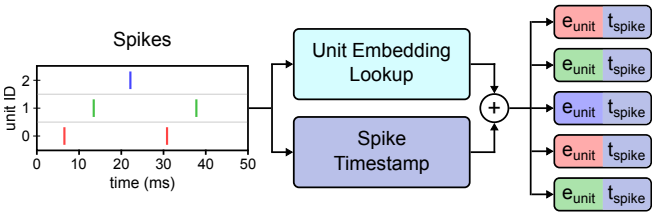

Figure 2: Individual spike tokenization scheme adapted from Azabou et al. (2023).

Input Cross-Attention. Similar to POYO, we use cross-attention as an encoder to build latent representations from individual spike tokens, where queries are from a trainable set of latent tokens and key-value pairs from input spikes. However, here we did so in an iterative fashion. As the encoder iterates through all available time intervals, the latents cross-attend with the spike tokens from the same interval alone.

State-Space Model and Output Projection. The result of the input cross-attention is then sent as an input to a Mamba model (Gu & Dao, 2024) that tracks a state across these short context windows. The hidden states from the Mamba model are then used as key-value pairs in an output-cross attention module that can be queried at different timestamps within the interval. Each query has a session embedding that captures differences due to experimental setup and other session-specific attributes that influence the recordings.

Results

We conducted experiments on two publicly available motor BCI datasets of nonhuman primates (Perich, Miller, Azabou, & Dyer, 2025; O'Doherty, Cardoso, Makin, & Sabes, 2020), comprising 151 recording sessions spanning five individuals performing either centre-out (CO) or random target (RT) reaching. The goal of decoding is to predict the 2D cursor velocities given recorded neural activity. The data was divided into batches such that input sequences to the model were all 1 second in length. Following Azabou et al. (2023), 14 sessions were held out of the multi-session training and used to evaluate the model's generalization to new sessions and animals.

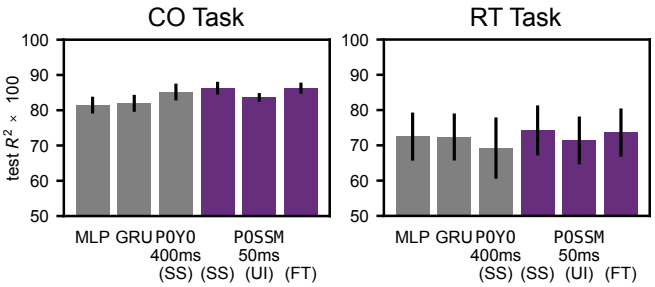

Figure 3: Decoding accuracy on centre-out (CO; left) and random target (RT; right) reaching tasks for a held-out subject. POSSM variants include single-session models (SS), multi-session models fine-tuned with just unit identification (UI), and multi-session models that are fully fine-tuned (FT).

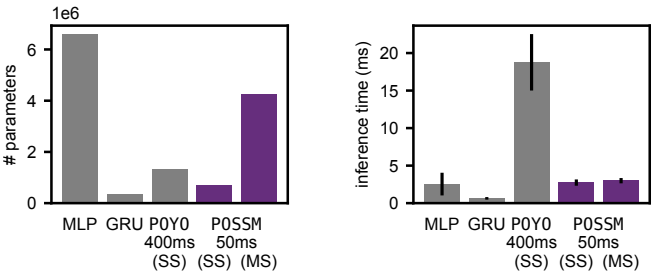

Figure 4: Comparing inference time per prediction (left) and model size (right). POSSM variants are single-session (SS) and multi-session (MS) models. POSSM is faster and more lightweight than POYO.

In addition to single-session and multi-session POSSM, various baseline methods were tested, including gated recurrent unit (GRU) (Cho et al., 2014), multi-layer perceptron (MLP) (Glaser et al., 2020), and POYO (Azabou et al., 2023) with a context length of 400ms. Our results are presented in Fig. 3 and 4.

Conclusion

We introduce a novel Transformer-SSM hybrid architecture that allows for real-time neural decoding and easy generalization to other datasets. We find that the performance of this architecture is comparable to or better than state-of-the-art approaches, while remaining lightweight and with low inference latency. With larger multi-dataset models, we show the benefits of data scaling for effective generalization to new sessions and individuals. In all cases, our models were at most half the size of a comparable POYO model (605K vs 1.3M for single-session, 4M vs 13M for multi-session). In the future, we wish to use self-supervised learning to learn meaningful representations across more datasets and even species. With effective pre-training, we could eventually build a foundation model for neural decoding, which could be fine-tuned to different several BCI tasks including motor, speech (Willett et al., 2023), and handwriting decoding (Willett, Avansino, Hochberg, Henderson, & Shenoy, 2021).

Acknowledgments

The authors would like to thank Mehdi Azabou, Eva Dyer, Patrick Mineault, and Blake Richards for support and feedback. MGP acknowledges support of a Future Leaders award from the Brain Canada Foundation and a J1 Chercheurs-boursiers en intelligence artificielle from the Fonds de recherche du Québec – Santé. GL acknowledges support from NSERC Discovery Grant RGPIN-2018-04821, the Canada Research Chair in Neural Computations and Interfacing, a Canada CIFAR AI Chair, the Digital Research Alliance of Canada, as well as IVADO and the Canada First Research Excellence Fund. The authors would also like to thank Mila and NVIDIA for providing computational resources that enabled this research.

References

Azabou, M., Arora, V., Ganesh, V., Mao, X., Nachimuthu, S., Mendelson, M., ... Dyer, E. (2023). A unified, scalable framework for neural population decoding. In *Advances in neural information processing systems* (Vol. 36, pp. 44937–44956). Curran Associates, Inc.

Chandravadia, N., Liang, D., Schjetnan, A. G. P., Carlson, A., Faraut, M., Chung, J. M., ... Rutishauser, U. (2020, March). A nwb-based dataset and processing pipeline of human single-neuron activity during a declarative memory task. *Scientific Data*, *7*(1). doi: 10.1038/s41597-020-0415-9

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014, October). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1724–1734). Doha, Qatar: Association for Computational Linguistics. doi: 10.3115/v1/D14-1179

Collinger, J. L., Wodlinger, B., Downey, J. E., Wang, W., Tyler-Kabara, E. C., Weber, D. J., ... Schwartz, A. B. (2013, February). High-performance neuroprosthetic control by an individual with tetraplegia. *The Lancet*, *381*(9866), 557–564. doi: 10.1016/s0140-6736(12)61816-9

Flesher, S. N., Downey, J. E., Weiss, J. M., Hughes, C. L., Herrera, A. J., Tyler-Kabara, E. C., ... Gaunt, R. A. (2021). A brain-computer interface that evokes tactile sensations improves robotic arm control. *Science*, *372*(6544), 831–836. doi: 10.1126/science.abd0380

Glaser, J. I., Benjamin, A. S., Chowdhury, R. H., Perich, M. G., Miller, L. E., & Kording, K. P. (2020). Machine learning for neural decoding. *eNeuro*, *7*(4). doi: 10.1523/ENEURO.0506-19.2020

Gu, A., & Dao, T. (2024). *Mamba: Linear-time sequence modeling with selective state spaces.*

Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., ... Carreira, J. (2022). Perceiver IO: A general architecture for structured inputs & outputs. In *International conference on learning representations.*

O'Doherty, J. E., Cardoso, M. M. B., Makin, J. G., & Sabes, P. N. (2020, May). *Nonhuman primate reaching with multichannel sensorimotor cortex electrophysiology.* Zenodo. doi: 10.5281/zenodo.3854034

Perich, M. G., Miller, L. E., Azabou, M., & Dyer, E. L. (2025). *Long-term recordings of motor and premotor cortical spiking activity during reaching in monkeys.* DANDI Archive.

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., & Liu, Y. (2023). *Roformer: Enhanced transformer with rotary position embedding.*

Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M., & Shenoy, K. V. (2021, May). High-performance brain-to-text communication via handwriting. *Nature*, *593*(7858), 249–254. doi: 10.1038/s41586-021-03506-2

Willett, F. R., Kunz, E. M., Fan, C., Avansino, D. T., Wilson, G. H., Choi, E. Y., ... Henderson, J. M. (2023, August). A high-performance speech neuroprosthesis. *Nature*, *620*(7976), 1031–1036. doi: 10.1038/s41586-023-06377-x

Wu, W., Black, M., Gao, Y., Serruya, M., Shaikhouni, A., Donoghue, J., & Bienenstock, E. (2002). Neural decoding of cursor motion using a kalman filter. In *Advances in neural information processing systems* (Vol. 15). MIT Press.