RETENTION BASED AUTOREGRESSIVE MODELS FOR MODELLING NEURAL DYNAMICS

BY

ABHINAV MURALEEDHARAN

A thesis submitted in conformity with
the requirements for the degree of
Masters in Engineering
Graduate Department of Institute for Aerospace Studies
University of Toronto

ABSTRACT

Retention based autoregressive models for modelling neural dynamics

Abhinav Muraleedharan

Masters in Engineering

Graduate Department of Institute for Aerospace Studies

University of Toronto

2023

In this work, we present Retention, a novel autoregressive model for generative modelling of sequences. Unlike Transformer based autoregressive models, retention scales linearly with respect to context size. We apply retention based models for modelling neural dynamics and achieve SOTA performance in neural modelling and behaviour decoding.

To Mom,

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my thesis supervisor, Prof. Prasanth Nair, for his invaluable guidance, unwavering support, and endless knowledge throughout the completion of my master's thesis. Prof. Nair's insightful feedback and constructive criticism played a pivotal role in shaping the direction of my research. I thoroughly enjoyed our discussions on various ideas, and his mentorship greatly enriched my academic experience.

I would also like to extend my appreciation to my co-supervisor, Taufik Valiante, for his motivating presence and the engaging discussions we had during the course of this research. Mr. Valiante's expertise and enthusiasm for the subject matter inspired me to tackle challenges with a fresh perspective.

Furthermore, I am grateful to both Prof. Nair and Mr. Valiante for not only guiding me in my academic pursuits but also providing valuable career advice that will undoubtedly influence my future endeavors.

This acknowledgment would be incomplete without expressing my deepest gratitude to my parents. Their unwavering support, encouragement, and belief in my abilities have been the cornerstone of my academic journey. Their sacrifices and love have fueled my determination, and I am profoundly thankful for their presence in my life.

Thank you all for your invaluable contributions and support.

PUBLICATIONS

This work has not been published yet.

CONTENTS

1 INTRODUCTION 1

2 PROBLEM STATEMENT 3

3 METHODS 5
- 3.1 Motivation . . . 5
- 3.2 Retention . . . 6
- 3.3 Training Retention Based Models . . . 7
- 3.4 Architecture . . . 7

4 RESULTS 8
- 4.1 A New Section . . . 8
- 4.2 Another Section in This Chapter . . . 9
- 4.3 Some Math . . . 10

5 CONCLUSION 11
- 5.1 A New Section . . . 11
- 5.2 Another Section in This Chapter . . . 12
- 5.3 Some Math . . . 13

1 INTRODUCTION

Hard problems inspire the creation of novel algorithms. These novel algorithms then find application in various contexts, distant from the original application which it was designed for. Understanding the human brain, specifically its dynamics and the relationship between dynamics of the brain and behaviour is a hard problem. Unraveling the dynamics of the brain holds the key to understanding the neural mechanisms underlying these processes. Beyond modeling neural activity, elucidating how such activity correlates with an organism's behavior is crucial for developing Brain-Computer Interfaces, clinical treatments for conditions like epilepsy, depression, and other neurodegenerative diseases. In this thesis, we develop scalable methods for modelling neural dynamics and relationship between the dynamics of the brain and behaviour. Although methods in this thesis are developed specifically for neural data, we believe that our approach would find application in diverse sequence modelling tasks in language, finance and engineering.

Machine learning techniques have played a pivotal role in modeling brain dynamics, and modeling the correlation between neural dynamics and behavior of an animal. In [POC+18], Pandarinath et al. introduced LFADS, an RNN-based method to infer latent dynamics from neural data. More recently, transformer-based models [VSP+17, GZ22], have been applied to learn neural dynamics and behaviour model. In [YP21], Pandarinath et al applied transformer-based models to learn neural dynamics without an explicit dynamical model. While LFADS and NDT (Neural Data Transformers) were focused on learning neural dynamics from single trial recordings, Azabou et.al recently introduced POYO [AAG+23], a transformer-based model to learn neural dynamics from multi-session neural recordings.

Although transformer-based models have shown remarkable success in general language modelling tasks and more recently in learning population dynamics of neurons, they exhibit poor scaling properties especially when applied to neural spiking data. Furthermore, unlike text data, neural recording probes sample on the order of kHz, and hence are characterized by high temporal resolution. This unique temporal aspect of neural

spiking data presents a challenge for transformers, which are originally designed for sequential data but may struggle with the high-frequency nature of neural signals. The transformer's poor scaling properties become particularly evident when recording from a large number of neurons simultaneously, as the number of potential firing patterns exponentially increases with the number of neurons.

In this work, we introduce a new class of autoregressive models to overcome limitations imposed by the architecture of attention-based transformer models. Our model has an unbounded context length and hence can capture long-range dependencies in the time series dataset. Furthermore, the complexity of training and inference of the parametrized model is independent of the context length, and hence our approach is computationally more efficient when compared to transformer-based autoregressive models.

2

PROBLEM STATEMENT

Imagine we are recording data from D neurons distributed across different regions of the brain. Let $x(t_i) \in \mathbb{R}^D$ denote the observed neural activity at timestep t_i and let y_i denote the observed behaviour of the animal at timestep t_i. From the time series dataset $\mathcal{D} = \{(x_i, y_i, t_i)\}_{i=1}^N$ of neural recordings, our goal is to construct:

- A predictive model of underlying brain dynamics
- A probabilistic model to predict behaviour of the organism at time $t+1$ given brain recordings until timestep t.

More formally, let's assume that the spiking activity is generated by an underlying non-stationary stochastic process defined by $p_t(x)$.

$$x(t) \sim p_t(x) \tag{2.1}$$

The probability of observing a sequence of neural recordings and behavior can be expressed as:

$$p(\{x_1, y_1\}, \{x_2, y_2\}, \{x_3, y_3\}, ..) = \lim_{N \to \infty} \prod_{i=1}^{N} p(\{x_i, y_i\} | \{x_1, y_1\}, \{x_2, y_2\}, ..\{x_{i-1}, y_{i-1}\}) \tag{2.2}$$

In the context of neural recordings, it is convenient to assume that the neural recording data and behavior can be modelled with separate probability distributions of the form:

$$\prod_{i=1}^{N} p_d(\{x_i\} | \{x_1\}, \{x_2\}, ..\{x_{i-1}\}) \tag{2.3}$$

$$\prod_{i=1}^{N} p_b(\{y_i\} | \{x_1\}, \{x_2\}, ..\{x_{i-1}\}) \tag{2.4}$$

Specifically, we assume that that the neural observed neural spiking data at timestep t_i is not dependent on the behavior variables in the preceding timesteps. Probability distributions of this nature have been extensively investigated in the field of language modeling. In conventional autoregressive frameworks, the approximation of conditional distributions often involves the utilization of parameterized models constrained by a finite

context limit [VSP+17]. While autoregressive models of this kind have been extremely successful in generating plausible language [RNS+18], they still struggle to capture long-range dependencies due to the finite context length limit[Hah20]. Furthermore, the complexity of training and inference of transformer-based models is $\mathcal{O}(N^2)$, where N is the context length of the transformer model.

3

METHODS

In this chapter, we describe the theory behind Retention based autoregressive models and methods for training retention based models.

3.1 MOTIVATION

To model conditional distributions defined in eq(2.4) exactly, we require a method that can process sequences of variable input length. In the realm of neural spiking data, which is frequently recorded at a sampling rate expressed in kHz, we require a method that can efficiently scale with the length of the sequence Furthermore, the patterns of neural activity increase exponentially with respect to the number of neurons, and hence the number of discrete tokens required to represent neural spike pattern at time step t can be prohibitively large. For instance, if we are recording spike signals from 100 neurons, in total there are 2^{100} possible firing patterns. Existing Transformer based models cannot be directly applied to model conditional distributions of these kind, without placing an assumption on the nature of probability distribution.

To address these challenges, we introduce "Retention", a mathematical operation to map a sequence of vectors $\{x_i\}_{i=1}^{N}$ to real valued vector ζ_i of same dimension. Retention is inspired from Score-life programming [Mur23], a novel method to solve sequential decision making problems. In Score-life programming [Mur23], Muraleedharan et.al applied the insight that the binary expansion of a real number can be used to represent a sequence of discrete variables. After constructing the mapping between a sequence of discrete variables and real numbers in a bounded interval, functions can be directly defined on the real numbers. By defining functions using this approach, we can model non-trivial relationships between elements of a sequence. In prior work [Mur23], has showed that such functions have unique properties, which can be exploited in developing efficient methods for solving determinsitic reinforcement learning problems. In our work, we extend this insight to vector valued variables, which are typically encountered in deep learning settings.

3.2 RETENTION

Mathematically, retention is defined as an exponentially weighted sum of a sequence of discrete vectors. If the vectors are drawn from a continuous space, then we perform thresholding operation to discretize the vectors. Specifically, given a sequence of vectors $\{x_i\}_{i=1}^{N}$, $x_i \in \mathbb{R}^d$, Retention variable $\zeta_k \in [0,1)^d$ as:

$$\zeta_k = \sum_{j=1}^{k} 2^{-(\log M)j} \sum_{i=0}^{M-1} \sigma(w_i \otimes x_{k-j+1} + b_i) \tag{3.1}$$

Here, $w_i, b_i \in \mathbb{R}^d$ are trainable parameters for the thresholding operation defined in inner summation. Given a vector $x_i \in \mathbb{R}^d$ as input, the inner summation operation acts like a smoothened step function, essentially discretizing elements of the vector to discrete values in the set: $\{0, 1, 2, , , M-1\}$. The outer summation operation, with an exponentially decaying factor maps the sequence of discrete vectors to a continuous real valued vector ζ_k. If the input data is discrete, or binary as in the case of neural spike signals, then the thresholding operation can be omitted, and the retention variable can be defined as:

$$\zeta_i = \sum_{k=1}^{i-1} 2^{-k}(x_{i-k}) \tag{3.2}$$

Given a sequence of discrete vectors $\{x_i\}_{i=1}^{N}$, retention variable ζ_i stores the discrete vectors in the binary expansion of ζ_i. If the vectors $\{x_i\}_{i=1}^{N}$ are continuous, then we perform a thresholding operation first to discretize the vectors and perform discounted sum of these discretized vectors.

Retention can also be defined for sequence of matrices $\{\mathbf{X}_i\}_{i=1}^{N}$ as:

$$\mathbf{1}_k = \sum_{j=1}^{k} 2^{-(\log M)j} \sum_{i=0}^{M-1} \sigma(\mathbf{W}_i \otimes \mathbf{X}_{k-j+1} + \mathbf{B}_i) \tag{3.3}$$

Modelling Conditional Distributions with Retention Variables

Now, we approximate the conditional distribution defined in eq(3) with retention. Specifically, the product of conditional distributions can now be approximated as:

$$\prod_{i=1}^{N} p_d(\{x_i\}|\{x_1\}, \{x_2\}, ..\{x_{i-1}\}) \approx \prod_{i=1}^{N} p_d(\{x_i\}|\zeta_i) \tag{3.4}$$

In this case, we

To learn the dynamics of the brain from neural recordings in an unsupervised manner, we maximize the following likelihood:

$$\mathcal{L}(X, \theta) = \sum_i log(p_d(\{x_i\}|\zeta_i; \theta)) \tag{3.5}$$

Here, $X = \{x_1, x_2, ...x_M\}$, the dataset of neural recordings.

Note that in this approach, the context window is not bounded, and the complexity of learning the parametrized model $p_d(\{x_i\}|\zeta_i; \theta)$ is independent of the length of the context window. While training the model, we apply eq(6) to recursively update ζ_i in an online fashion, instead of pre-computing and storing $\{\zeta_i\}_{i=1}^N$ separately.

To learn the correlation between neural dynamics and behavior, we follow a similar approach and approximate the conditional distribution defined in eq(4) with:

$$\prod_{i=1}^{N} p_b(\{y_i\}|\{x_1\}, \{x_2\}, ..\{x_{i-1}\}) \approx \prod_{i=1}^{N} p_b(\{y_i\}|\zeta_i) \tag{3.6}$$

We define the loss function associated with this approach as the negative log-likelihood of the observed behavioral outcomes given the estimated neural activity states. Formally, the loss function $\mathcal{L}$ is expressed as:

$$\mathcal{L}(X, Y, \phi) = -\sum_i \log p_b(\{y_i\}|\zeta_i; \phi)$$

3.3 TRAINING RETENTION BASED MODELS

In this section, we describe how retention based models are trained.

Online Computation of Retention Variable

Offline Computation of Retention Variable

3.4 ARCHITECTURE

Two Photon Dataset: U-Net

Fully connected networks

4

RESULTS

Ei choro aeterno antiopam mea, labitur bonorum pri no [Gau27]. His no decore nemore graecis. Suavitate interpretaris eu, vix eu libris efficiantur.

4.1 A NEW SECTION

Illo principalmente su nos. Non message *occidental* angloromanic da. Debitas effortio simplificate sia se, auxiliar summarios da que, se avantiate publicationes via. Pan in terra summarios, capital interlingua se que. Al via multo esser specimen, campo responder que da. Le usate medical addresses pro, europa origine sanctificate nos se.

Examples: *Italics*, ALL CAPS, SMALL CAPS, LOW SMALL CAPS.

Colours: red, green, blue

Test for a Subsection

Note: The content of this chapter is just some dummy text. It is not a real language.

Lorem ipsum at nusquam appellantur his, ut eos erant homero concludaturque. Albucius appellantur deterruisset id eam, vivendum partiendo dissentiet ei ius. Vis melius facilisis ea, sea id convenire referrentur, takimata adolescens ex duo. Ei harum argumentum per. Eam vidit exerci appetere ad, ut vel zzril intellegam interpretaris.

Errem omnium ea per, pro con populo ornatus cu, ex qui dicant nemore melius. No pri diam iriure euismod. Graecis eleifend appellantur quo id. Id corpora inimicus nam, facer nonummy ne pro, kasd repudiandae ei mei. Mea menandri mediocrem dissentiet cu, ex nominati imperdiet nec, sea odio duis vocent ei. Tempor everti appareat cu ius, ridens audiam an qui, aliquid admodum conceptam ne qui. Vis ea melius nostrum, mel alienum euripidis eu.

Ei choro aeterno antiopam mea, labitur bonorum pri no. His no decore nemore graecis. In eos meis nominavi, liber soluta vim cu.

Autem Timeam

Nulla fastidii ea ius, exerci suscipit instructior te nam, in ullum postulant quo. Congue quaestio philosophia his at, sea odio autem vulputate ex.

Cu usu mucius iisque voluptua. Sit maiorum propriae at, ea cum primis intellegat. Hinc cotidieque reprehendunt eu nec. Autem timeam deleniti usu id, in nec nibh altera.

4.2 ANOTHER SECTION IN THIS CHAPTER

Non vices medical da. Se qui peano distinguer demonstrate, personas internet in nos. Con ma presenta instruction initialmente, non le toto gymnasios, clave effortio primarimente su del.[1]

Sia ma sine svedese americas. Asia representantes un nos, un altere membros qui.[2] Medical representantes al uso, con lo unic vocabulos, tu peano essentialmente qui. Lo malo laborava anteriormente uso.

(a) Illo secundo continentes sia il, sia russo distinguer se. Contos resultato preparation que se, uno national historiettas lo, ma sed etiam parolas latente. Ma unic quales sia. Pan in patre altere summario, le pro latino resultato.

(b) Lo vista ample programma pro, uno europee addresses ma, abstracte intention al pan. Nos duce infra publicava le. Es que historia encyclopedia, sed terra celos avantiate in. Su pro effortio, o.

Tu uno veni americano sanctificate. Pan e union linguistic del le, del un apprende denomination.

Personas Initialmente

Uno pote summario methodicamente al, uso debe nomina hereditage ma. Iala rapide ha del, ma nos esser parlar. Maximo dictionario sed al.

A Subsubsection

Deler utilitate methodicamente con se. Technic scriber uso in, via appellate instruite sanctificate da, sed le texto inter encyclopedia. Ha iste americas que, qui ma tempore capital.

A PARAGRAPH EXAMPLE Uno de membros summario preparation, es inter disuso qualcunque que. Del hodie philologos occidental al, como publicate litteratura in web. Veni americano es con, non internet millennios secundarimente ha.

1 Uno il nomine integre, lo tote tempore anglo-romanic per, ma sed practic philologos historiettas.

2 De web nostre historia angloromanic.

Figure 4.1: Some smart caption

Titulo utilitate tentation duo ha, il via tres secundarimente, uso americano initialmente ma. De duo deler personas initialmente. Se duce facite westeuropee web, nos clave articulos ha.

Medio integre lo per, non es linguas integre. Al web altere integre periodicos, in nos hodie basate. Uno es rapide tentation, usos human synonymo con ma, parola extrahite Figure 5.1 greco-latin ma web. Veni signo rapide nos da.

4.3 SOME MATH

We can define scalar multiplication in $\mathbb{R}^n$ by

$$c[u_1, \ldots, u_n] = [cu_1, \ldots, cu_n]$$

You can now check that for $u, v \in \mathbb{R}^n$, we have

$$c(u + v) = cu + cv$$

5 CONCLUSION

Ei choro aeterno antiopam mea, labitur bonorum pri no [Gau27]. His no decore nemore graecis. Suavitate interpretaris eu, vix eu libris efficiantur.

5.1 A NEW SECTION

Illo principalmente su nos. Non message *occidental* angloromanic da. Debitas effortio simplificate sia se, auxiliar summarios da que, se avantiate publicationes via. Pan in terra summarios, capital interlingua se que. Al via multo esser specimen, campo responder que da. Le usate medical addresses pro, europa origine sanctificate nos se.

Examples: *Italics*, ALL CAPS, Small Caps, low small caps.

Colours: red, green, blue

Test for a Subsection

Note: The content of this chapter is just some dummy text. It is not a real language.

Lorem ipsum at nusquam appellantur his, ut eos erant homero concludaturque. Albucius appellantur deterruisset id eam, vivendum partiendo dissentiet ei ius. Vis melius facilisis ea, sea id convenire referrentur, takimata adolescens ex duo. Ei harum argumentum per. Eam vidit exerci appetere ad, ut vel zzril intellegam interpretaris.

Errem omnium ea per, pro con populo ornatus cu, ex qui dicant nemore melius. No pri diam iriure euismod. Graecis eleifend appellantur quo id. Id corpora inimicus nam, facer nonummy ne pro, kasd repudiandae ei mei. Mea menandri mediocrem dissentiet cu, ex nominati imperdiet nec, sea odio duis vocent ei. Tempor everti appareat cu ius, ridens audiam an qui, aliquid admodum conceptam ne qui. Vis ea melius nostrum, mel alienum euripidis eu.

Ei choro aeterno antiopam mea, labitur bonorum pri no. His no decore nemore graecis. In eos meis nominavi, liber soluta vim cu.

Autem Timeam

Nulla fastidii ea ius, exerci suscipit instructior te nam, in ullum postulant quo. Congue quaestio philosophia his at, sea odio autem vulputate ex.

Cu usu mucius iisque voluptua. Sit maiorum propriae at, ea cum primis intellegat. Hinc cotidieque reprehendunt eu nec. Autem timeam deleniti usu id, in nec nibh altera.

5.2 ANOTHER SECTION IN THIS CHAPTER

Non vices medical da. Se qui peano distinguer demonstrate, personas internet in nos. Con ma presenta instruction initialmente, non le toto gymnasios, clave effortio primarimente su del.[1]

Sia ma sine svedese americas. Asia representantes un nos, un altere membros qui.[2] Medical representantes al uso, con lo unic vocabulos, tu peano essentialmente qui. Lo malo laborava anteriormente uso.

(a) Illo secundo continentes sia il, sia russo distinguer se. Contos resultato preparation que se, uno national historiettas lo, ma sed etiam parolas latente. Ma unic quales sia. Pan in patre altere summario, le pro latino resultato.

(b) Lo vista ample programma pro, uno europee addresses ma, abstracte intention al pan. Nos duce infra publicava le. Es que historia encyclopedia, sed terra celos avantiate in. Su pro effortio, o.

Tu uno veni americano sanctificate. Pan e union linguistic del le, del un apprende denomination.

Personas Initialmente

Uno pote summario methodicamente al, uso debe nomina hereditage ma. Iala rapide ha del, ma nos esser parlar. Maximo dictionario sed al.

A Subsubsection

Deler utilitate methodicamente con se. Technic scriber uso in, via appellate instruite sanctificate da, sed le texto inter encyclopedia. Ha iste americas que, qui ma tempore capital.

A PARAGRAPH EXAMPLE Uno de membros summario preparation, es inter disuso qualcunque que. Del hodie philologos occidental al, como publicate litteratura in web. Veni americano es con, non internet millennios secundarimente ha.

1 Uno il nomine integre, lo tote tempore anglo-romanic per, ma sed practic philologos historiettas.

2 De web nostre historia angloromanic.

Figure 5.1: Some smart caption

Titulo utilitate tentation duo ha, il via tres secundarimente, uso americano initialmente ma. De duo deler personas initialmente. Se duce facite westeuropee web, nos clave articulos ha.

Medio integre lo per, non es linguas integre. Al web altere integre periodicos, in nos hodie basate. Uno es rapide tentation, usos human synonymo con ma, parola extrahite Figure 5.1 greco-latin ma web. Veni signo rapide nos da.

5.3 SOME MATH

We can define scalar multiplication in $\mathbb{R}^n$ by

$$c[u_1, \ldots, u_n] = [cu_1, \ldots, cu_n]$$

You can now check that for $u, v \in \mathbb{R}^n$, we have

$$c(u + v) = cu + cv$$

BIBLIOGRAPHY

[AAG+23] Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael J Mendelson, Blake Richards, Matthew G Perich, Guillaume Lajoie, and Eva L Dyer, *A unified, scalable framework for neural population decoding*, arXiv preprint arXiv:2310.16046 (2023).

[Gau27] Carl Friedrich Gauß, *General investigations of curved surfaces*, 1827.

[GZ22] Nicholas Geneva and Nicholas Zabaras, *Transformers for modeling physical systems*, Neural Networks **146** (2022), 272–289.

[Hah20] Michael Hahn, *Theoretical limitations of self-attention in neural sequence models*, Transactions of the Association for Computational Linguistics **8** (2020), 156–171.

[Mur23] Abhinav Muraleedharan, *Beyond dynamic programming*, arXiv preprint arXiv:2306.15029 (2023).

[POC+18] Chethan Pandarinath, Daniel J O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky, Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg, et al., *Inferring single-trial neural population dynamics using sequential auto-encoders*, Nature methods **15** (2018), no. 10, 805–815.

[RNS+18] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al., *Improving language understanding by generative pre-training*.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, *Attention is all you need*, Advances in neural information processing systems **30** (2017).

[YP21] Joel Ye and Chethan Pandarinath, *Representation learning for neural population activity with neural data transformers*, arXiv preprint arXiv:2108.01210 (2021).

COLOPHON

This thesis was typeset using the typographical look-and-feel `classicthesis` developed by André Miede and Ivo Pletikosić.

The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*".

Here you can insert things like "Figures were created with..."

[Insert version number/description, if you want]