Predicting human behavioral decisions with recurrent neural networks

Yu-Ang Cheng* (yuang_cheng@brown.edu)

Ivan Felipe Rodriguez* (ivan_felipe_rodriguez@brown.edu)

Thomas Serre (thomas_serre@brown.edu)
Department of Cognitive Linguistic & Psychological Sciences
Brown University
Providence, RI, USA

Abstract

Current neural network modeling work in visual recognition has focused primarily on matching behavioral choices and related accuracy measures. Visual perception is a dynamic process that unfolds in time, but moving beyond characterizing choice patterns to capturing temporal aspects of visual decision-making has been challenging. We introduce a novel computational framework to optimize recurrent neural networks (RNNs) response times. First, we consider a random dot motion task and show how an RNN can be fitted to human psychophysics data. Second, we train an ideal observer RNN model to maximize a tradeoff between speed and accuracy. Our results indicate that human-like reaction time distributions can naturally emerge in a neural network explicitly optimized to solve a task in minimal computing time. Finally, we use our approach with a biological-plausible circuit model of decision-making known as the Wong-Wang model (Wong & Wang, 2006). We show that it is possible to stack this module on top of a task-optimized convolutional neural network to fit human behavioral data. Overall, our results suggest that the proposed framework can be effectively used to fit models of visual perception with the full set of human behavioral data, bringing us one step closer to an integrated model of human visual perception.

Keywords: recurrent neural network; human-AI alignment; reaction times

Introduction

Neural networks have been widely used to model the visual system with recent neural architectures achieving near human-level accuracy on complex visual categorization tasks (Serre, 2019). Engineering considerations have largely driven the development of computational models towards scaling them for better performance. However, a growing body of literature highlights increasing misalignment between modern deep neural networks and primate vision (Fel et al., 2022; Linsley et al., 2023). Critically, current deep neural networks lack the ability to account for reaction time. In psychological research, reaction time is critical because it provides insights into cognitive functions' processing speed and efficiency (Heitz, 2014). It has been widely known that such processing speed has a complex relationship with accuracy measures (Heitz, 2014). Therefore, considering reaction times is essential for developing a complete understanding of human vision (Mnih et al., 2014; Lake et al., 2017).Recurrent Neural Networks (RNNs) constitute a promising framework to model the visual system as these models can be used to integrate feedforward processes with feedback ones (Kar et al., 2019; Wyatte et al., 2014). Furthermore, RNNs explicitly incorporate a notion of time via recurrence steps, which can be used as a proxy for reaction time. Unfortunately, time in an RNN is not directly differentiable, and hence, standard back-propagation cannot be used out of the box to fit RNNs to human reaction times. As a result, researchers have resorted to surrogate metrics such as measures of uncertainty or confidence levels derived from machine learning to approximate reaction times (Goetschalckx et al., 2023). Here, we develop a mathematical approximation of reaction time for RNNs that can be tuned via back-propagation. This novel framework involves setting adaptive thresholds within the network across time. Thus, it can potentially approximate human reaction times better than alternative approaches. Additionally, we extend this framework to convolutional neural networks (CNNs), applying a similar approach to modify their architecture with a biologically plausible circuit, the Wong-Wang model (Wong & Wang, 2006). The proposed method is versatile, and we show that it has the potential to enhance a wide range of neural network architectures, making significant strides toward the development of a complete model of human visual perception.

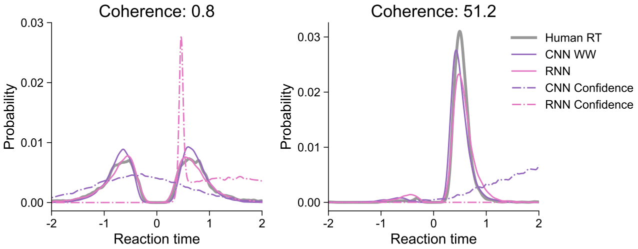

Figure 1: Reaction time distribution fitting. Using a confidence value cannot capture the full distribution of reaction times (RNN confidence & CNN confidence). Instead, dynamically thresholding an RNN output or stacking a canonical RNN (WW) onto a CNN output can fit reaction times well.

[1] These authors contributed equally to this work.

Differentiable reaction time framework

We begin by defining a recurrent neural network where the state at time step $t \in \{1,\ldots,N\}$ is denoted h_t. Let us assume that when the hidden state reaches a specific boundary at any time T, the network is compelled to make a decision, meaning that it cannot process inputs beyond time T. For simplicity, we assume the boundary to be linear in the space of the hidden state. Mathematically, let us define $T(h) = \min\{t \in \{1,\ldots,N\} : W_h h_t > 0\}$. Where W_h is a linear transformation of the hidden state at time t.

The primary challenge that arises when fitting the model's reaction time $T(h)$ with human reaction times is the non-differentiability of $T(h)$, which prevents the use of the backpropagation algorithm. This is because $T(h)$ is an integer that requires non-differentiable operations such as the minimum function and inequality. We approximate the differentiation using a first-order Taylor series expansion.

Let us redefine $T(\tilde{h(t)})$ as $\min\{t > 0 : \tilde{h(t)} > 0\}$, where $\tilde{h}(t) = W_h h(t)$. Considering a small perturbation in the hidden state using first-order Taylor expansions, we get:

$$T(\tilde{h} + \delta\tilde{h}) \approx T(\tilde{h}) + \frac{dT}{d\tilde{h}}\delta\tilde{h} \tag{1}$$

Since T is given by the time step in which $\tilde{h(t)}$ reaches the threshold, then the change $\tilde{h}(t + (\frac{d\tilde{h}}{dt})^{-1}\delta\tilde{h})$ introduces a proportional change for our defined T, because it means h_t would reach the activity threshold at a different time. Therefore

$$T(\tilde{h} + \delta\tilde{h}) \approx T(\tilde{h}) - (\tfrac{d\tilde{h}}{dt})^{-1}\delta\tilde{h} \tag{2}$$

Combining step 1 and step 2 we can reach the quantity:

$$\frac{dT}{d\tilde{h}} \approx -\left(\frac{d\tilde{h}}{dt}\right)^{-1} \approx \frac{1}{\tilde{h}(t-1) - \tilde{h}(t)}$$

This means that the function to reach the threshold of activity is found to be approximately equal to the inverse of the rate of change of the evolution of the activity $\tilde{h}$. In other words, if we take a piece-wise approximation of the activity between time steps, the time it takes to reach the threshold divided by the difference between the current state to the distance is the inverse of the rate change (speed).

Task & Stimuli

To validate our approach, we trained a Recurrent Neural Network (RNN) on the Random Dot Motion (RDM) task, a classic paradigm used extensively in psychophysics studies (Ball & Sekuler, 1987), human imaging (Shibata et al., 2012), and electrophysiology (Law & Gold, 2008). The stimuli in this task consist of dots moving on a screen with different levels of coherence toward a predefined direction vs. randomly. Our RNN is a 5-layered convolution network with a 4096-unit LSTM. The network was trained for 100 epochs using the Adam optimizer with a learning rate of 1e-4 for the first 10 epochs and 1e-5 for the rest. The results showed that it performed this task at a near human-like level, achieving nearly 100% accuracy under the high coherence conditions (e.g., 51.2%) and approximating chance level at low coherence (e.g., 0.8%).

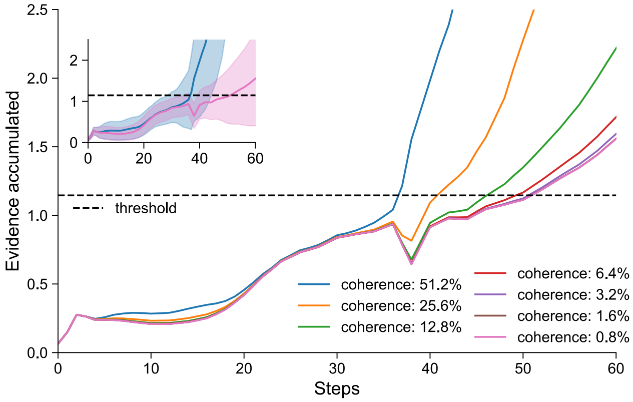

Figure 2: Evidence accumulation of the RNN model across time. For higher coherence levels mean reaction time is faster, for lower coherence levels mean reaction time is slower. The inset shows the standard deviation of reaction times, where for the highest coherence (blue), the reaction times are less varied compared with the lowest coherence (pink)

To explore the alignment between humans and the model, we keep the model parameters frozen and only train a linear boundary upon the RNN output to fit a signed reaction time distribution derived from human psychophysics (original experiment here (Green et al., 2010)) (Figure 1, Human RT). Surprisingly, our model closely matches the human reaction times (Figure 1, RNN). For comparison, we also adopt the previous method that fits a confidence value (a linear transformation of time-weighted logits) from the model to approximate reaction times.(Figure 1, RNN confidence), and it does not work.

A biological-plausible drop-in module for fitting human RT

The Wong-Wang model is a popular mechanistic neural circuit decision-making model (Wong & Wang, 2006). In its reduced form, this model resembles an Ornstein-Uhlenbeck process in which two competing populations interact. When a threshold is crossed, a decision is made in a winner-takes-all fashion. Traditionally, the input to the model is a steady value, proportional to the coherence level of the stimulus, e.g if the stimuli is generated at 51% coherence the model would get as input 0.51 at every single step. One significant limitation of the model is that it requires meticulous parameter hand tuning. Here, we offer two improvements to this model. First, we replace the constant input (a constant coherence level) with the output of a stimulus-computable neural network. Second, we make the model trainable allowing to fit the entire model to human reaction time data. Results are shown in Figure 1 (CNN WW).

Conclusions

In this work, we have introduced a trainable framework to train RNNs to learn to adjust a decision threshold so that decisions can be made dynamically based on a variable number of time steps. We showed that such optimization can be used to fit an RNN directly to human reaction times. We also showed that such a framework can be used to learn a stimulus-specific penalize the recurrence steps of RNNs, and human-like reaction times of the RNN can naturally emerge even when no human data is provided. Finally, we have transformed a popular RNN for decision-making, the Wong Wang model, into a trainable module that can be stacked into any neural network (RNNs and CNNs) enabling any model to fit human reaction times.

Acknowledgments

This work was supported by ONR (N00014-24-1-2026).

References

Ball, K., & Sekuler, R. (1987). Direction-specific improvement in motion discrimination. *Vision research*, *27*(6), 953–965.

Fel, T., Rodriguez, I. F., Linsley, D., & Serre, T. (2022). Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in Neural Information Processing Systems (NeurIPS)*.

Goetschalckx, L., Govindarajan, L. N., Karkada Ashok, A., Ahuja, A., Sheinberg, D., & Serre, T. (2023). Computing a human-like reaction time metric from stable recurrent vision models. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 14338–14365). Curran Associates, Inc.

Green, C. S., Pouget, A., & Bavelier, D. (2010). Improved probabilistic inference as a general learning mechanism with action video games. *Current biology*, *20*(17), 1573–1579.

Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, 150.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019, April). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.*.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, E253.

Law, C.-T., & Gold, J. I. (2008). Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area. *Nature neuroscience*, *11*(4), 505–513.

Linsley, D., Rodriguez, I. F., Fel, T., Arcaro, M., Sharma, S., Livingstone, M., & Serre, T. (2023). *Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex.*

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., . . . Hassabis, D. (2014). Recurrent models of visual attention. In *Advances in neural information processing systems* (Vol. 27).

Serre, T. (2019, September). Deep learning: The good, the bad, and the ugly. *Annu Rev Vis Sci*, *5*, 399–426.

Shibata, K., Chang, L.-H., Kim, D., Náñez Sr, J. E., Kamitani, Y., Watanabe, T., & Sasaki, Y. (2012). Decoding reveals plasticity in v3a as a result of motion perceptual learning.

Wong, K.-F., & Wang, X.-J. (2006, January). A recurrent network mechanism of time integration in perceptual decisions. *J. Neurosci.*, *26*(4), 1314–1328.

Wyatte, D., Jilk, D. J., & O'Reilly, R. C. (2014, July). Early recurrent feedback facilitates visual object recognition under challenging conditions. *Front. Psychol.*, *5*, 674.