RTify: Aligning Deep Neural Networks with Human Behavioral Decisions

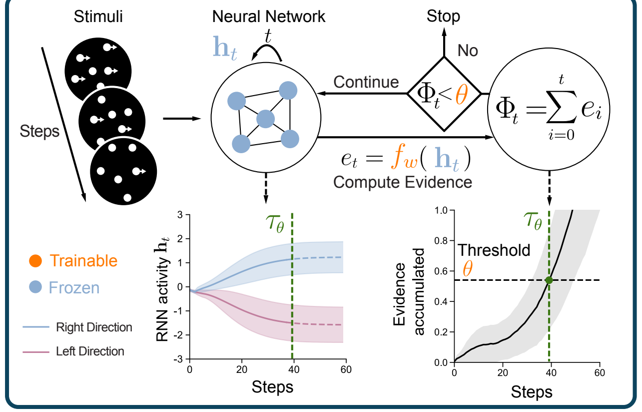
Yu-Ang Cheng*[1], Ivan Felipe Rodriguez*[1], Sixuan Chen[1], Kohitij Kar[2], Takeo Watanabe[1], Thomas Serre[1]

1 Department of Cognitive & Psychological Sciences, Carney Center for Computational Brain Science, Brown University

2 Department of Biology, York University

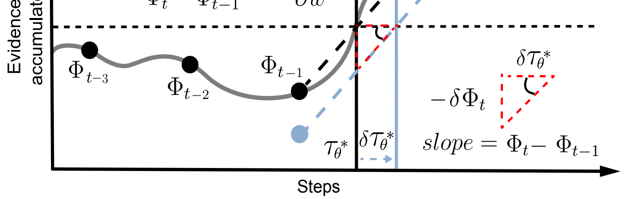
Read the Paper Here

- Current **neural network models** of primates primarily replicate **behavioral accuracy** but overlook **reaction times (RTs),** a key indicator of **visual perception dynamics**.
- **Decision-making models** (e.g. DDM, LBA) have focused on explaining how visual information gets integrated over time, **but cannot handle more complex, natural stimuli** apart from parameterized stimuli (e.g. Gabor).
- **Recurrent neural networks (RNNs)** hold great promise since they **are temporally dynamic, image-computable, and have a notion of RT** via recurrence steps.



$$\frac{\partial L(\tau_\theta)}{\partial w} = \frac{\partial L(\tau_\theta(\Phi))}{\partial \tau_\theta(\Phi)} \cdot \frac{\partial \tau_\theta(\Phi)}{\partial w} \approx \frac{\partial L(\tau_\theta(\Phi))}{\partial \tau_\theta(\Phi)} \cdot \frac{\partial \tau^*_\theta(\Phi)}{\partial w}$$

$$\frac{\partial \tau^*_\theta}{\partial w} \approx \frac{\delta \tau^*_\theta}{\delta w} \approx -\frac{\delta \Phi_t(\frac{1}{\Phi_t - \Phi_{t-1}})}{\delta w} \approx \frac{1}{\Phi_t - \Phi_{t-1}} \cdot \left(-\frac{\partial \Phi_t}{\partial w}\right)$$



- We present RTify, a novel computational approach to optimize the recurrent steps of RNNs to account for human RTs.
- With this framework, we **successfully fit an RNN directly to human behavioral responses**.
- Our framework can also be extended to an ideal-observer model whereby the RNN is trained without human data via a penalty term that **encourages the network to make a decision as quickly as possible.**
- Under this setting, **human-like behavioral responses naturally emerge from the RNN.**

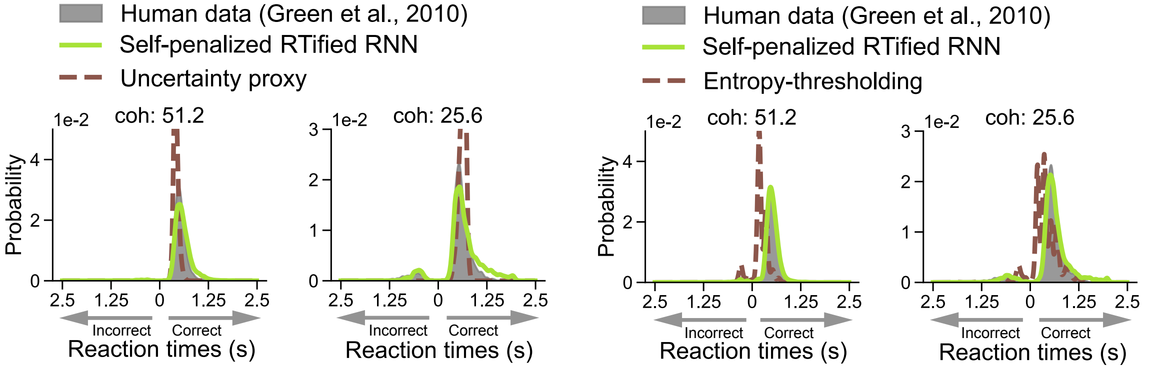
Extracting RTs with supervision (fitting human RTs)

$$Loss = \|\mathcal{D}_{\text{human RT}} - \mathcal{D}_{\text{model RT}}\|_2$$

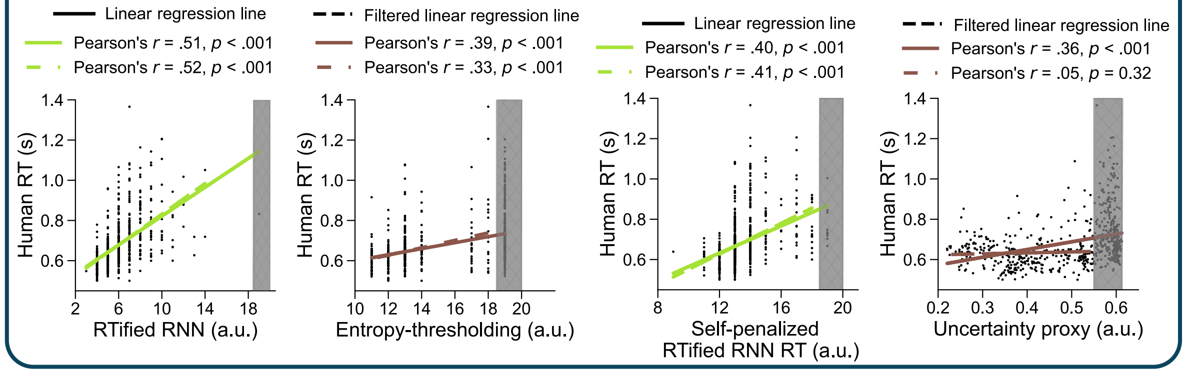
Extracting RTs via self penalty (leveraging RNN dynamics)

$$Loss = CCE + \lambda \cdot l_y \cdot \mathcal{T}_\tau$$

Model Evaluation on Random Dot Motion Task



Model Evaluation on Object Recognition Task



Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., & Kriegeskorte, N. (2020). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. PLoS computational biology, 16(10), e1008215.
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.
Kar, K., Kubilius, J., Schmidt, K., Issa, E.B., DiCarlo, J.J.: Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. Nature neuroscience 22(6) (2019) 974–983。
Goetschalckx, L., Govindarajan, L. N., Karkada Ashok, A., Ahuja, A., Sheinberg, D., & Serre, T. (2024). Computing a human-like reaction time metric from stable recurrent vision models. Advances in Neural Information Processing Systems, 36.
Acknowledgment: This work was supported by NSF (IIS-2402875), ONR (N00014-24-1-2026) and the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANR-19-PI3A-0004) to T.S and National Institutes of Health (NIH R01EY019466 and R01EY027841) to T.W. Computing hardware was supported by NIH Office of the Director grant (S10OD025181) via Brown's Center for Computation and Visualization (CCV).