Reverse-engineering Recurrent Neural Network solutions to a hierarchical inference task for mice

Rylan Schaeffer[1,2], Mikail Khona[3], Leenoy Meshulam[2,4], International Brain Laboratory[2], Ila Rani Fiete[2,4,5]

[1]Harvard Institute for Applied Computational Science [2]International Brain Laboratory [3]MIT Department of Physics [4]MIT Department of Brain and Cognitive Sciences [5]MIT McGovern Institute for Brain Research

NEURAL INFORMATION PROCESSING SYSTEMS

Research Aims

Goal: Reverse engineer how recurrent neural networks perform hierarchical inference involving two latent variables and disparate time scales separated by 1-2 orders of magnitude.

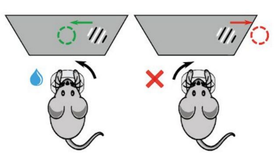
Research Questions:

1. How well do RNNs compare against normative Bayesian baselines?
2. What are the representations, dynamics and mechanisms RNNs use to perform inference?

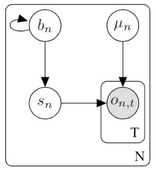
Research Contributions:

- Answers to the above questions.
- Propose a novel, task-agnostic distillation technique for extracting interpretable circuits from high dimensional task-trained models.

Hierarchical Inference Task



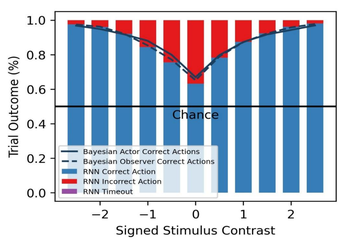
$$b_n \sim P(b|b_{n-1}, p_b(l_n))$$
$$s_n = b_n | b_n \sim Bern(p_s)$$
$$\mu_n \sim \mathcal{U}([0, 0.5, 1.0, 1.5, 2.0, 2.5])$$
$$o^S_{n,t} | \mu_n \sim \mathcal{N}(\mu_n, 1)$$
$$o^{\sim S}_{n,t} | \mu_n \sim \mathcal{N}(0, 1)$$

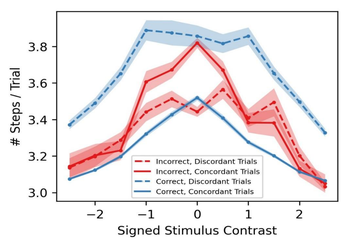


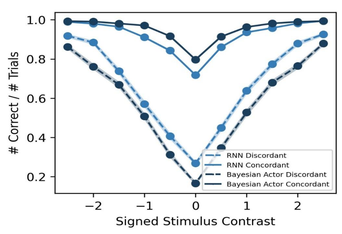
Each trial is a two-alternative forced choice visual stimulus task. Over a number of consecutive trials (a block), the stimulus has a higher probability of appearing on one side; in the next block, the stimulus side probabilities switch. Both stimulus side and block side must be inferred. RNNs are trained via gradient descent on cross entropy summed over all steps across multiple blocks; the target distribution is the stimulus side in each trial.

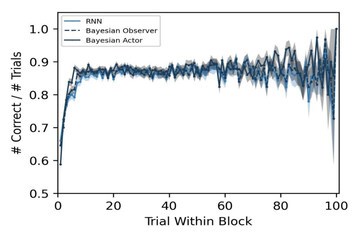
Behavior

RNNs quantitatively match normative Bayesian models. Psychometric curves show near-optimal inference and change-point detection.



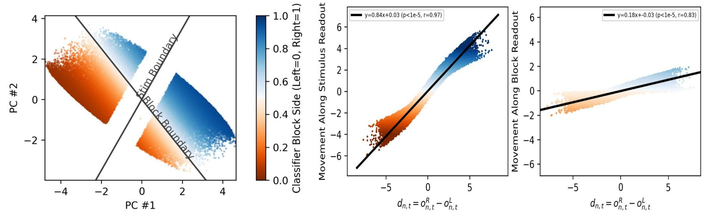




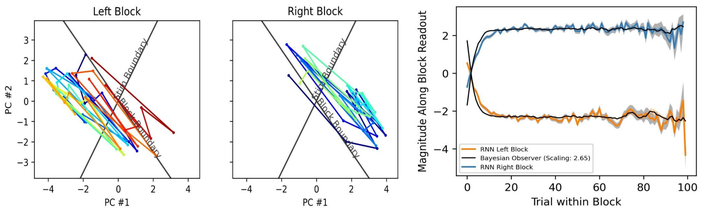


Representations

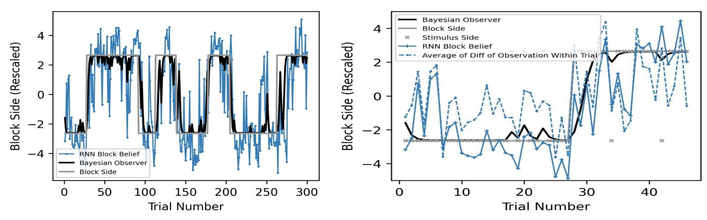
RNNs learn 2-dimensional encoding of stimulus side and block side. Sensory evidence at each time step pushes the RNN state along both the stimulus direction and the block direction.



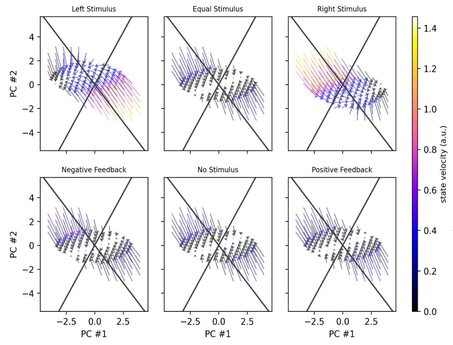
Movement along both directions is integrated across RNN steps to infer the stimulus side and block side. The RNN state evolves quickly over the stimulus side decision boundary, on a trial by trial basis, and evolves slowly along the block side direction.



The RNN's dynamics induce coupling between the two inference processes, causing the RNN to perform below Bayes optimality.



Dynamics

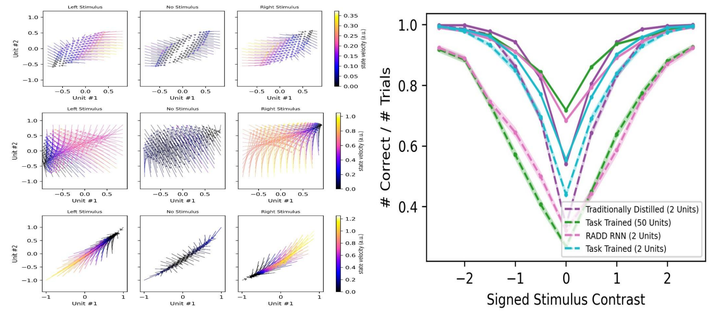


The RNN displays two kinds of dynamical behavior:

1. Within-trial flows to one of two discrete fixed points, driven by sensory evidence.
2. Between-trial relaxation to line attractor, maintaining estimate of block side.

Mechanism

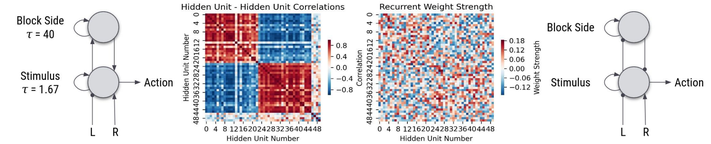
To extract the effective circuit, we propose a novel distillation technique: Representations and Dynamics Distillation (RADD). A 2-unit RADD RNN (top) reproduces the phase portraits (left) and psychometric curves (right) of high-dim task-trained RNNs, whereas a 2-unit knowledge distilled RNN [1,2,3] (middle) and a 2-unit task-trained RNN (bottom) do not.



The 2-unit RADD RNN yields interpretable, highly sensible parameters that show recurrent interactions to support integration via different amplitudes.

$$\hat{z}_{n,t} = \begin{bmatrix} \text{Stim Belief}_{n,t} \\ \text{Block Belief}_{n,t} \end{bmatrix} = \tanh\left(\begin{bmatrix} 0.54 & 0.31 \\ 0.19 & 0.84 \end{bmatrix} \hat{z}_{n,t-1} + \begin{bmatrix} -0.20 & 0.20 & 0.005 \\ -0.04 & 0.04 & 0.02 \end{bmatrix} \begin{bmatrix} o^L_{n,t} \\ o^R_{n,t} \\ r_{n,t} \end{bmatrix} + \begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix} \right)$$

The 2-unit RADD RNN circuit (left) has sensible integration timescales, and we find that the high-dim, task-trained RNNs learn equivalent circuits (right).



References

[1] Cristian Bucilă, Rich Caruana, and Alexandru Niculescu-Mizil. "Model compression". Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006, pp. 535–541.

[2] Jimmy Ba and Rich Caruana. "Do Deep Nets Really Need to be Deep?" Advances in Neural Information Processing Systems 27. 2014, pp. 2654–2662.

[3] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. "Distilling the Knowledge in a Neural Network". In:arXiv:1503.02531 [cs, stat].

Acknowledgements

INSTITUTE FOR APPLIED COMPUTATIONAL SCIENCE AT HARVARD UNIVERSITY

MCGOVERN INSTITUTE FOR BRAIN RESEARCH

INTERNATIONAL BRAIN LABORATORY

brain+cognitive sciences