Overcoming sensory-memory interference in artificial and biological neural networks

Andrii Zahorodnii (zaho@mit.edu)
Massachusetts Institute of Technology, Cambridge, MA, USA

Diego Mendoza-Halliday (diegomendoza@pitt.edu)
Massachusetts Institute of Technology, Cambridge, MA, USA
University of Pittsburgh, Pittsburgh, PA, USA

Ning Qian (nq6@columbia.edu)
Columbia University, New York, NY, USA

Robert Desimone (desimone@mit.edu)
Massachusetts Institute of Technology, Cambridge, MA, USA

Christopher J. Cueva (ccueva@gmail.com)
Massachusetts Institute of Technology, Cambridge, MA, USA

Memories of recent stimuli are crucial for guiding behavior. However, the same sensory pathways that receive information to be remembered are constantly bombarded by new sensory experiences, and it remains largely unknown how the brain overcomes interference between sensory and memory representations. Here we report which mechanisms might be at play in artificial and biological networks that are robust to sensory-memory interference. We examined recurrent neural networks (RNNs) that were either hand-designed or trained using gradient descent methods, and compared our results with neural data from two macaque experiments. We found an infinite RNN solution space, that included gating of the sensory inputs, modulating synapse strengths to achieve a strong attractor solution, and dynamic coding of feature preference, such that, at the extreme, cells invert their tuning over time. Neural data from macaque brain area medial superior temporal (MST) was most aligned with the Gating + Inversion of Tuning solution. This solution was also consistent with experimental results from monkey behavior. Taken together, our results help elucidate how recurrent neural networks are able to solve the problem of sensory-memory interference using a combination of both static and dynamic codes, and suggest MST may play a role in this computation.

Keywords: working memory; attractor models; recurrent neural networks; sensory-memory interference; cortex

Our behavior is guided both by immediate sensory experiences and by memories of recently encountered stimuli (Figure 1a). Many studies explore how neural circuits store memories. However, it remains largely unknown how these memory systems both allow information to flow into them while also preserving this information as we continue to interact with the world (Libby & Buschman, 2021; Cueva* et al., 2021).

The problem of sensory-memory interference is likely widespread as biologically relevant variables are often encoded by broadly tuned cells, for example, direction and orientation tuning of MT and V1 neurons (Schiller et al., 1976; Albright, 1984). If two stimulus orientations are presented successively at the same location (S. Ding et al., 2017), they provide similar inputs, via the same set of connections, to the same set of memory units. How, then, does the system prevent the memory of the first orientation from being overwritten by the arrival of the second? Outside of controlled experimental settings, the problem of sensory-memory interference must still be overcome by neural circuits as eye movements realign relevant stimuli so we effectively have sequential presentations of stimuli in the same retinotopic positions, much like the experimental settings.

Results

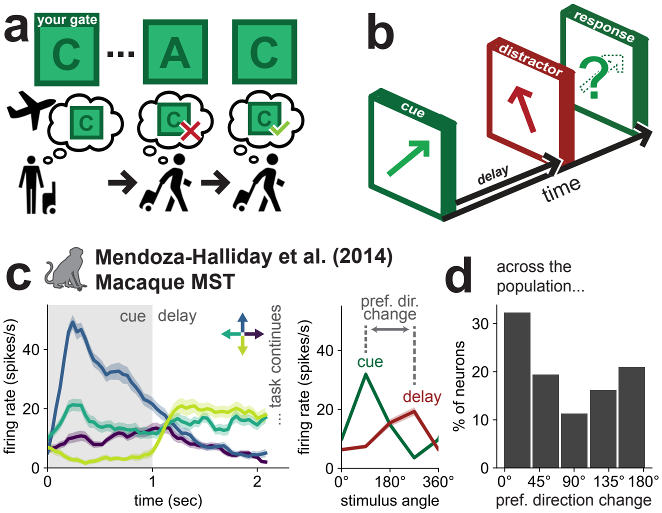

Figure 1: (a) Sensory-memory interference in the real world, at the airport. We see our gate assignment, C, and then can keep this in memory even as we see other gates on our search through the airport. (b) The distractor task. A cue is shown and after a variable time interval a distractor is shown. The goal of this task is to produce a response at the initial input direction while ignoring the distractor. (c) Neurons from the macaque brain area MST invert their tuning preferences over time between the cue and delay periods. The firing rate of a single neuron is shown here for four cue directions, along with tuning curves during cue and delay periods. (d) Preference changes for the entire population of MST neurons.

To ground our study of sensory-memory interference with experimental data, we model a working memory task that includes a distractor stimulus as described in Figure 1b (Suzuki & Gottlieb, 2013). The goal of this task is to remember the first stimulus and ignore a subsequently presented distractor. To better understand the underlying computations required to overcome the problem of sensory-memory interference we trained, examined, and eventually were able to hand-design recurrent neural networks (RNNs) to solve this problem (Figure 2a-d). The dynamics of the simulated neurons were governed by the standard continuous-time RNN equation that has previously been used to model neural responses (Mante* et al., 2013; Sussillo et al., 2015).

We looked for ways of experimentally differentiating between these solutions and found that in the delay period between the cue and distractor, our models made distinct predictions for the dynamics of neurons' preferred directions (Figure 2e-h). Notably, some networks used a dynamic coding scheme such that cells changed their tuning to prevent the new sensory input from overwriting the previously stored one.

To see if any of these potential mechanisms might be employed by the brain, we analyzed neural recordings from two experiments on nonhuman primates (Mendoza-Halliday et al., 2014, 2024). In particular, we focused on area medial superior temporal (MST) because MST is uniquely situated between sensory and memory regions, and is the first area along the dorsal pathway to show sustained working memory activity (Mendoza-Halliday et al., 2014). MST may be uniquely positioned to confront the problem of sensory-memory interference, and intriguingly displays the same dynamic tuning patterns as some of our models (Figure 1c-d). Notably, our hypothesis is that even in tasks that do not explicitly contain a distractor, the brain may still "hide" memories in anticipation of future inputs.

We compared the firing rates of the models during the cue and delay periods to the two neural datasets from MST using the Procrustes distance metric (Williams et al., 2021; F. Ding et al., 2021) after concatenating the average activity from the cue and delay periods for every value of the cued stimulus. We found that one solution to the problem of sensory-memory interference is near the noise floor for one of the datasets, and is the best model for the other dataset as well (Figure 3a-c). This model also has behavioral signatures consistent with the monkey behavior reported by Suzuki & Gottlieb (2013), in contrast to a standard ring attractor model that strongly alters memories of recent stimuli with subsequent inputs (Figure 3d-e). Taken together, our results help elucidate how recurrent neural networks are able to solve the problem of sensory-memory interference by leveraging both static and dynamic codes, and bridges scales from behavior to neural firing patterns to synaptic connectivity. Intriguingly, our work also suggests that, even beyond the specific context of sensory-memory interference, the dynamic neural codes seen in the brain may enable information to effectively "hide" from being overwritten. Finally, we propose a new role for area MST in overcoming sensory-memory interference.

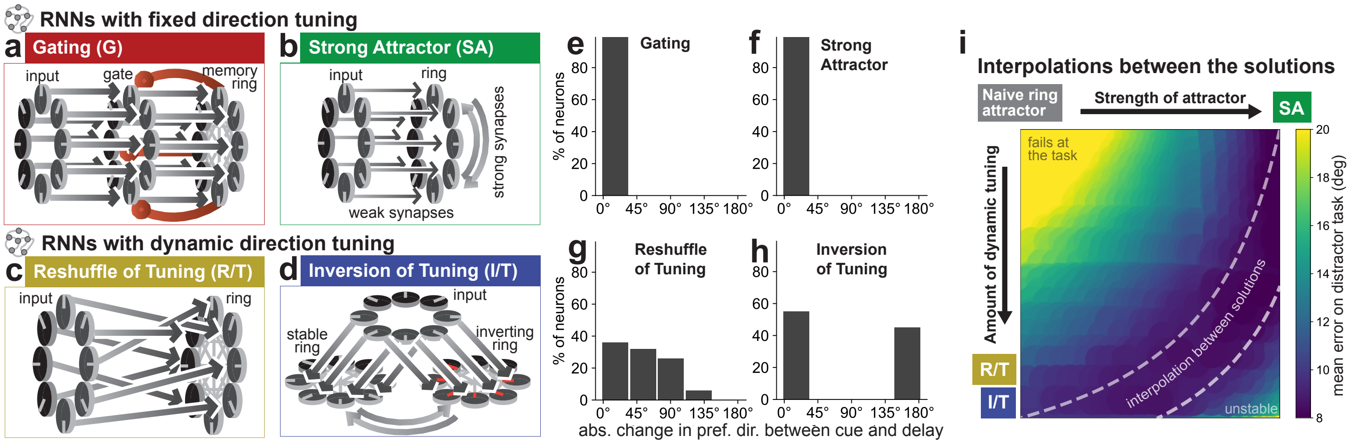

Figure 2: Different solutions lead to distinct experimental predictions about the connectivity between neurons, and how the preferred directions change between the time of the initial cue and the subsequent delay period. Hand-designed RNNs implementing (a) Gating, (b) Strong Attractor, (c) Reshuffle of Tuning, and (d) Inversion of Tuning. (e-h) Histogram of neurons' absolute preferred direction changes, as predicted by the corresponding model. (i) The sensory-memory interference problem can be overcome by interpolating between solution mechanisms.

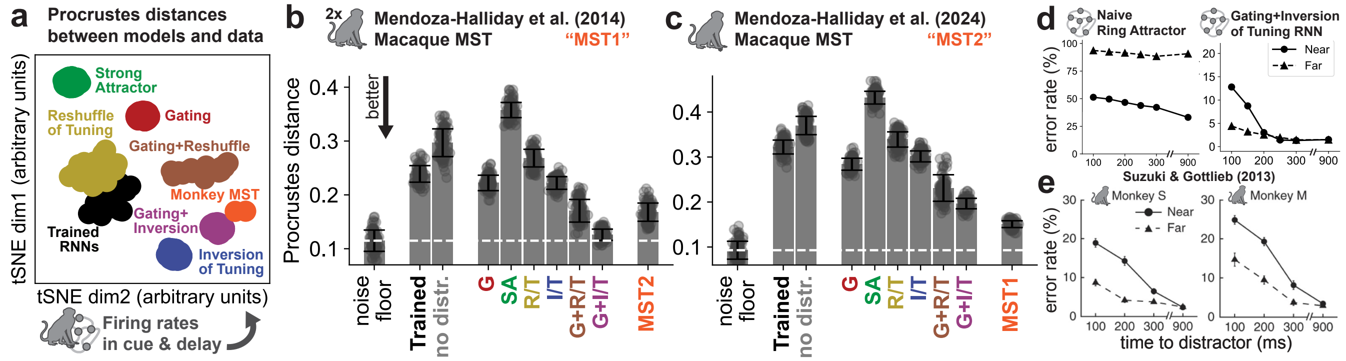

Figure 3: (a) The solution space of models and data is visualized using the t-SNE algorithm on the firing rates of neurons in the cue and delay periods. Every point represents a model network or neural dataset. (b, c) Procrustes distance between the models and neural responses from macaque brain area MST (Mendoza-Halliday et al., 2014, 2024). Across both experiments, neural data from MST is most aligned with the Gating + Inversion of Tuning mechanism. Error bars indicate the standard deviation. (d) Behavioral data (error rates) on the distractor task for a standard ring attractor network, which is not robust to the distractor, and the Gating + Inversion of Tuning RNN, which is similar to monkey behavior. To compare with monkey behavior from Suzuki & Gottlieb (2013) the distractor can either be similar (near) to the initial cue stimulus (45°away) or far (135°/180°away). (e) Corresponding behavioral results from two monkeys (Suzuki & Gottlieb, 2013).

References

Albright, T. D. (1984). Direction and orientation selectivity of neurons in visual area mt of the macaque. *Journal of Neurophysiology*, *52*(6), 1106-1130.

Cueva*, C. J., Ardalan*, A., Tsodyks, M., & Qian, N. (2021). Recurrent neural network models for working memory of continuous variables: activity manifolds, connectivity patterns, and dynamic codes. *arXiv:2111.01275*.

Ding, F., Denain, J.-S., & Steinhardt, J. (2021). Grounding representation similarity through statistical testing. In *Advances in neural information processing systems* (Vol. 34).

Ding, S., Cueva, C. J., Tsodyks, M., & Qian, N. (2017). Visual perception as retrospective bayesian decoding from high- to low-level features. *PNAS*, *114*(43), E9115–E9124.

Libby, A., & Buschman, T. J. (2021). Rotational dynamics reduce interference between sensory and memory representations. *Nature Neuroscience*, *24*, 715-726.

Mante*, V., Sussillo*, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*, 78-84.

Mendoza-Halliday, D., Torres, S., & Martinez-Trujillo, J. C. (2014). Sharp emergence of feature-selective sustained activity along the dorsal visual pathway. *Nature Neuroscience*, *17*, 1255-1262.

Mendoza-Halliday, D., Xu, H., Azevedo, F. A., & Desimone, R. (2024). Dissociable neuronal substrates of visual feature attention and working memory. *Neuron*, *112*(5), 850-863.e6.

Schiller, P. H., Finlay, B. L., & Volman, S. F. (1976). Quantitative studies of single-cell properties in monkey striate cortex. ii. orientation specificity and ocular dominance. *Journal of neurophysiology*, *39*(6), 1320–1333.

Sussillo, D., Churchland, M. M., Kaufman, M. T., & Shenoy, K. V. (2015). A neural network that finds a naturalistic solution for the production of muscle activity. *Nature neuroscience*, *18*, 1025-1033.

Suzuki, M., & Gottlieb, J. (2013). Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. *Nature Neuroscience*, *16*, 98–104.

Williams, A. H., Kunz, E., Kornblith, S., & Linderman, S. W. (2021). Generalized shape metrics on neural representations. In *Advances in neural information processing systems* (Vol. 34).