Geometry of naturalistic object representations in models of working memory

Xiaoxuan Lei (xiaoxuan.lei@mail.mcgill.ca)
Department of Physiology, 3655 Promenade Sir-William-Osler
Monteral, Quebec, Canada

Takuya Ito (taku.ito1@gmail.com)
T.J. Watson Research Center, IBM Research
Yorktown Heights, New York, United States

Pouya Bashivan (pouya.bashivan@mcgill.ca)
Department of Physiology, 3655 Promenade Sir-William-Osler
Monteral, Quebec, Canada

Abstract

Working memory (WM) is a central cognitive ability crucial for intelligent decision-making. Recent experimental and computational work studying WM has primarily been carried out using categorical stimuli (Panichello & Buschman, 2021; Yang et al., 2019), rather than ecologically-valid, multidimensional naturalistic inputs. Moreover, such studies have primarily evaluated WM on single or limited numbers of tasks. As a result, there is a lack of understanding in how naturalistic object information is processed by neural circuits. To bridge this gap, we developed sensory-cognitive models, consisting of a convolutional neural network (CNN) coupled with a recurrent neural network (RNN), and trained them on nine distinct N-back tasks using naturalistic stimuli. By examining the RNN's latent space, we found that: 1) multi-task RNNs simultaneously represent both task-relevant and irrelevant information while performing tasks; 2) the latent subspaces used to maintain specific object properties are largely stable across tasks in vanilla RNNs but not in gated ones; and 3) RNNs embed objects in new representational spaces in which individual object feature axes are more orthogonalized compared to the perceptual space, enhancing separation of features. Our findings elucidate the ways in which goal-driven RNNs adapt their latent representations in response to task requirements.

Keywords: working memory; recurrent neural networks; multi-tasking; representation analysis

Methods

Behavioral task suite. We considered N-back tasks ($N \in \{1,2,3\}$) based on one of three distinct object properties (i.e. feature; $F \in \{Location, Identity, Category\}$ (denoted as L, I, C), resulting in a total of 9 N-back task variants (Fig. 1a). Naturalistic stimuli were generated using 3D object models from the ShapeNet dataset (Klabunde et al., 2023), comprising 4 object categories, each with 2 unique identities rendered from various view angles, and presented at 1 of 4 possible locations. The training and validation datasets differed in their viewing angles, necessitating view-invariant processing by the model.

Model architecture. We developed a two-stage model that delineates perceptual and cognitive processes (Fig. 1b). This model processes sequences of images, utilizing an ImageNet pre-trained ResNet50 (He et al., 2016) model to derive visual embeddings. These embeddings, combined with a vector representation of tasks (i.e. task index), are fed into a discrete-time RNN module, which produces one of three possible responses: match, non-match, or no action at each time step. Each network is trained to perform one (single-task) or multiple tasks (multi-feature or multi-task or both). After training, we analyzed activations from the penultimate layer of ResNet50 (i.e. the perceptual space), as well as the RNN activations during the stimulus presentation and subsequent timesteps (i.e. encoding and memory space respectively). We considered the vanilla RNN, GRU, and LSTM.

Model training. We trained three groups of models that differed on their training diet: **1) single-task, single-feature**: trained on a single n-back task based on a single object feature (e.g.1-back location); **2) single-task, multi-feature**: trained on a single choice of *N* for all three feature variations (e.g. 1-back L, or C or I); **3) multi-task, multi-feature**: encompassing all choices of N (1,2,3) and features (L, I, C). All models reached $> 90\%$ accuracy on both train and validation datasets. The ensuing analyses utilized data collected from models with 512 units for vanilla RNNs, and 256 units for GRUs and LSTMs.

Results

Naturalistic object information in task-optimized RNNs. We first investigated whether task-relevant and -irrelevant object features are retained by each model during task performance. To probe this, we fitted decoders to predict each object property from the RNN hidden state activity from the first timestep of each trial (e.g. $F = L_i$ vs. $F = L_{j \neq i}$ decoders, total 4 location decoders). We found that while task-irrelevant object features are not preserved in single-task single-feature models (Fig. 2b left), they are well-preserved (i.e. decoding accuracy $> 85\%$) in multi-feature and multi-task models (Fig. 2b middle and right). In other words, the RNNs retained a complete picture of the object representation in their latent space regardless of which object properties were required for

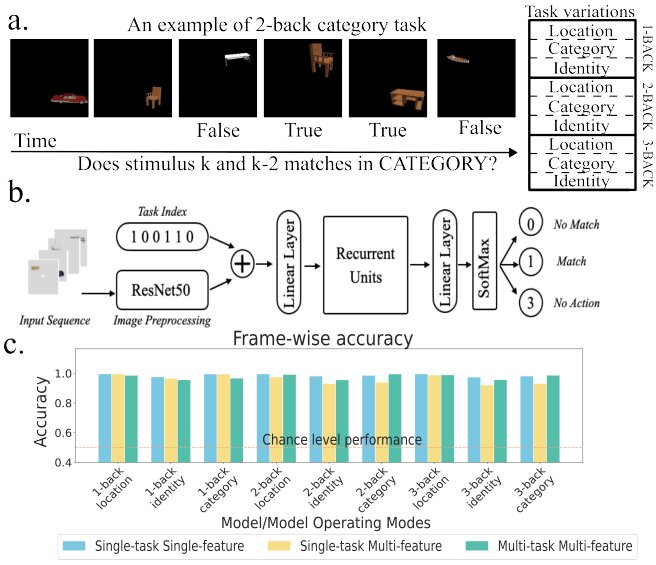

Figure 1: **Tasks and Models:** a) Left: An example of a 2-back category task; Right: 9 task variations of N-back constructed from different choices of task-relevant features (L, I, C) and N (1,2,3) index. b) The sensory-cognitive model architecture. c) Performance of GRU models.

performing the task [1].

Consistency of object representations in RNNs across tasks. Having observed that both task-relevant and -irrelevant information are retained by the multi-feature RNNs, we next asked whether information about object properties are reliably encoded in a common latent subspace within the RNN. To probe this, we trained decoders to predict object properties from the RNNs' activations, and then evaluated the decoder on other tasks (i.e., cross-task decoding). We found that the gated RNNs (GRU and LSTM) used highly task-specific subspaces to encode object properties, while vanilla RNNs encoded shared object properties across all task-variations (Fig.2a,c). This suggests that gated RNNs tend to learn task-specific subspaces that do not generalize across tasks, potentially impacting their ability to quickly adapt to new tasks. Moreover, these results provide a setup for testing hypotheses about the architecture of the recurrent mechanisms in brain areas such as prefrontal cortex underlying WM.

Representational orthogonalization in task-optimized RNNs. To improve their performance, RNN weights are likely optimized to obtain more structured and separable geometrical representation for each task-relevant feature. We hypothesized that to effectively solve the task, the RNN latent space may orthogonalize feature representations beyond their perceptual representation. To quantify orthogonalization, we calculated the angles between all pairs of decision hyperplanes using cosine similarity. We then summarized these angles into a single orthogonality measure by computing the Frobenius

[1] The observation was consistent across all tested model architectures unless otherwise specified.

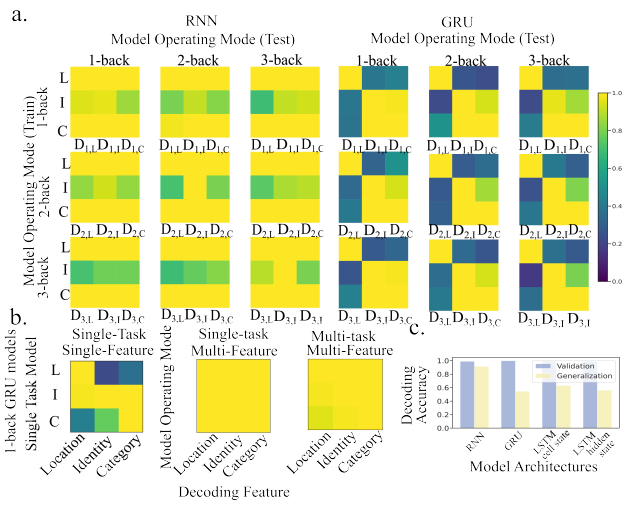

Figure 2: **Object representation in RNN hidden state: (a)** We trained decoders on task-relevant object properties and subsequently evaluated their generalization performance to other tasks. Each row and column of each big 3×3 heatmaps correspond to on which N-back task the decoders are trained and tested on. Within each heatmap, each column represents a different decoder, denoted by $D_{k,F}$ ($k \in \{1,2,3\}$, $F \in \{L,I,C\}$) (indicating which task and decoding feature the decoder was trained on), while each row corresponds to the object property of the task the decoder was tested on. Left: Vanilla RNN, Right: GRU. **(b)** Validation accuracy of decoders trained on RNN latent space activations from the first timestep of each trial to predict object properties. Each column in each heatmap represents the object property the decoder was trained on, while each row corresponds to a model. GRU models of Upper: single-task single-feature, Middle: single-task multi-feature, Bottom: multi-task multi-features. **c** Quantification of the validation accuracy (within the same task, indicated in purple) and generalization accuracy (across tasks with different task-relevant features, indicated in yellow) across all model architectures.

norm of the difference between this matrix and the identity matrix (denoting an idealized orthogonal space). We defined this measure as the orthogonalization index (O, Figure 3b). Consistent with our hypothesis, we observed that compared to the perceptual space, the RNN latent space orthogonalizes the axes along which distinct object features are represented, enabling structured and enhanced separation of these features.

Conclusion

By analyzing the representational geometry of different classes of RNNs, we showed that RNNs maintain object information regardless of their task relevance and they further orthogonalize the objects' embeddings beyond their perceptual representation. However, while the underlying subspace for encoding different object features are stable across tasks in vanilla RNNs, they are highly task-specific in gated RNNs.

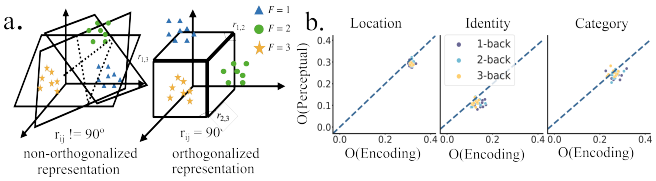

Figure 3: **Orthogonalization:** a) a schematic of two hypothetical object spaces in a 3D space. $r_{i,j}$ represents the angle formed by the decision hyperplanes that separate feature value i or j from the others. Left: non-orthogonalized representation; Right: orthogonalized representation. b) Normalized orthogonalization index, for both perceptual and encoding spaces respectively (denoted as $O(Perceptual)$ and $O(Encoding)$). Data points fall below the diagonal line indicate a more orthogonalized representation in the RNN encoding space.

Together, our analyses shed light on how high-dimensional object representations are maintained in multitask RNNs and highlights critical representational differences as a function of network architecture. Future work will test these predictions with human brain data during the performance of same tasks.

Acknowledgments

This research was supported by the Healthy-Brains-Healthy-Lives startup supplement grant and the NSERC Discovery grant RGPIN-2021-03035 (P.B.). X.L. was supported by UNIQUE Excellence Scholarship. P.B. was supported by FRQ-S Research Scholars Junior 1 grant 310924, and the William Dawson Scholar award.

References

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Klabunde, M., et al. (2023). Similarity of neural network models: A survey of functional and representational measures. *arXiv preprint arXiv:2305.06329*.

Panichello, M. F., & Buschman, T. J. (2021). Shared mechanisms underlie the control of working memory and attention. *Nature*, *592*(7855), 601–605. doi: 10.1038/s41586-021-03323-z

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., & Wang, X.-J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, *22*(2), 297–306. doi: 10.1038/s41593-018-0310-2