Computational cognitive science

https://doi.org/10.1038/s42256-024-00885-9

Cognitive maps from predictive vision

Margaret C. von Ebers & Xue-Xin Wei

Check for updates

Constructing spatial maps from sensory inputs is challenging in both neuroscience and artificial intelligence. A recent study demonstrates that a self-attention neural network using predictive coding can generate an environmental map in its latent space as an agent that navigates the environment.

A map of the environment helps us localize, plan and navigate in space. For decades, how to acquire spatial maps has been crucial for understanding the navigation system of the brain and for building efficient robotic systems. Neuroscience research has discovered functional cell types, including place cells[1] and grid cells[2], that are thought to support the cognitive map[3] — an internal representation of space in the brain. However, the formation of these representations is still debated. A study by Gornet and Thomson[4] in *Nature Machine Intelligence* reveals that learning to predict the flow of visual inputs may enable an agent to construct a spatial map.

In principle, inputs from multiple modalities (for example, visual, olfactory and vestibular) may be used to construct spatial maps. Prior research showed that spatial maps in the brain could be constructed by integrating the agent's movements from vestibular inputs, known as path integration[5]. This computation can be implemented in continuous attractor networks[6–8]. Recent work using deep learning showed that response patterns similar to grid cells emerged in recurrent neural networks optimized for path integration tasks[9–11]. While visual inputs have been recognized to play an important part in the anchoring and

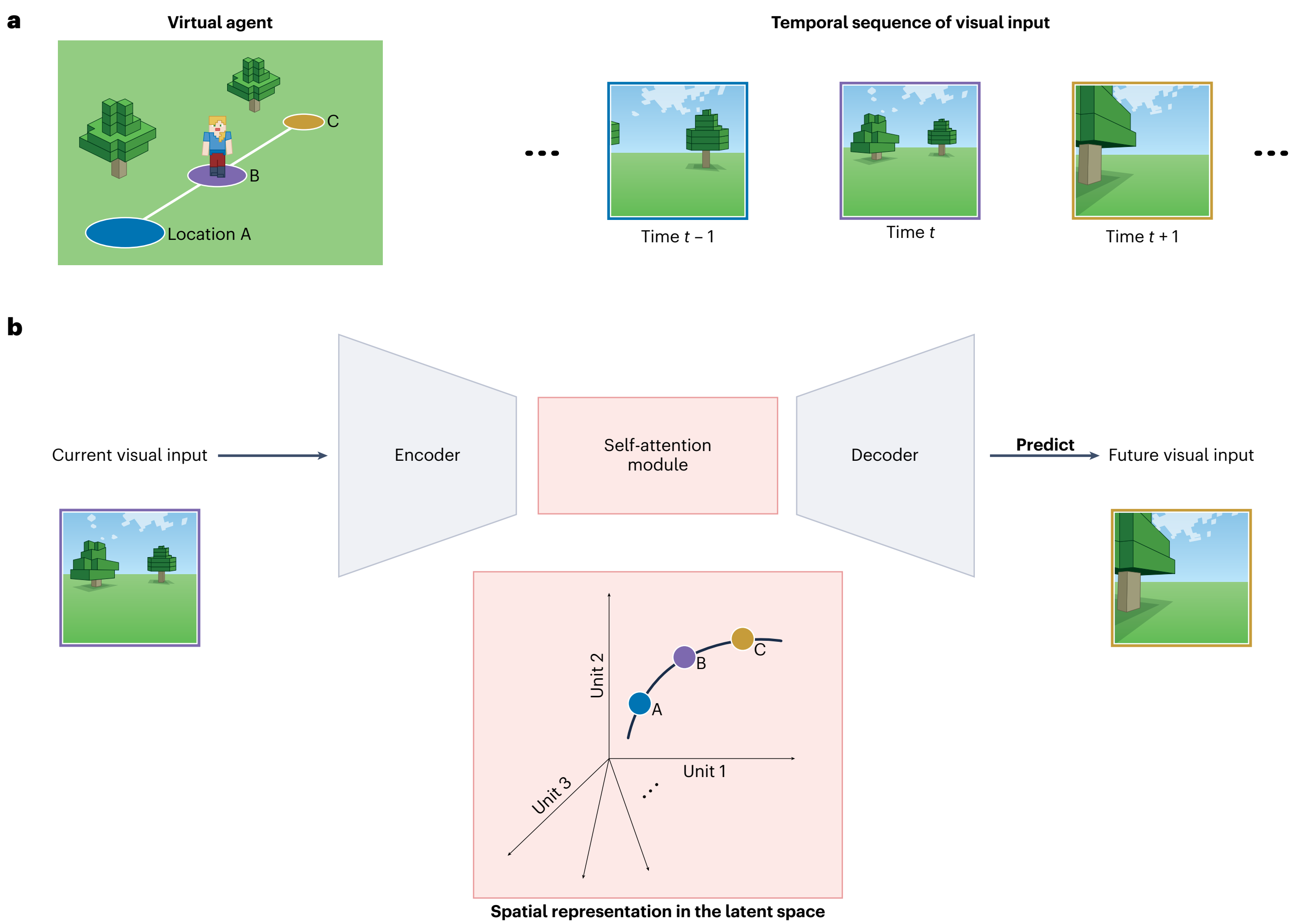

Fig. 1 | Visual input for a virtual agent and model architecture. a, A virtual agent navigates in an environment and collects sequences of visual input. These image sequences are used to train the model. **b**, Overview of the model architecture. The neural network was trained to predict future visual input on the basis of current visual input. A representation of space emerged from the self-attention module after training the network to solve the visual prediction task.

recalibration of the map, path integration is generally hypothesized to be a crucial component for map construction.

Inspired by predictive coding in computational neuroscience, Gornet and Thomson[4] explore a different hypothesis: learning spatial maps by predicting the future visual scene alone. Predictive coding proposes that the brain encodes a predictive representation by leveraging the spatial-temporal regularity of sensory inputs. Initially developed to explain inhibitory neural responses in the retina[12], seminal work by Rao and Ballard[13] extended this framework to explain puzzling response properties of neurons in the primary visual cortex. Gornet and Thomson[4] now apply this framework to spatial navigation.

To formalize their hypothesis, the authors developed a mathematical theory predicting that, to accurately predict future visual inputs, a navigational agent should (1) encode its spatial location and direction in the environment, and (2) learn the transition statistics of its movements. To test the theory, they trained a deep network model to predict future visual inputs for a simulated navigation agent. Their model is composed of several components (Fig. 1). First, an encoder converts each image in a sequence of temporally related observations into a latent vector. Second, a self-attention module analyses these latent vectors to generate a predicted latent vector for each time step. Third, a decoder, functioning as the inverse of the encoder, generates a prediction for the next frame's image. To generate training data, they collected sequences of observations from a Minecraft agent walking between random start and destination locations in an environment featuring trees, a cave landmark and a pond with a bridge. After training the agent to solve the visual prediction task, the authors investigated the latent representation in the output of the self-attention module.

They report that deep networks trained on this visual task can track the agent's location in the environment, decodable from network activity with considerable accuracy. Furthermore, individual model units developed spatial selectivity. This activity in the latent layer of the model captured the proximity structure of physical locations, aligning with the concept of a spatial map. This was supported by an analysis showing that distances in the latent neural activity space corresponded to physical distances (albeit with substantial variability), allowing for an approximate readout of the distances between various locations in the environment. The emergence of these spatial representations in the latent layers, based on a purely visual prediction task, corroborates their mathematical theory.

A question that arises is whether prediction is truly important for forming spatial maps. Simply reconstructing visual input frame-by-frame might be sufficient. To demonstrate the benefit of prediction, Gornet and Thomson[4] compared their model with a baseline model that reconstructs individual images without prediction. This baseline model should capture image similarity, which may or may not reflect the spatial relationship. The authors found that the predictive coding network developed a more accurate spatial map than the baseline model. They explored this question further by creating another environment, where the agent may encounter visually identical observations at different points in a circular hallway. Decoding the location of these aliased observations using the baseline model resulted in large prediction errors, while the predictive coder could correctly distinguish them.

Together, these results raise the intriguing possibility that agents solving prediction-based visual tasks may be sufficient to develop spatial representations, without relying on inputs encoding body movements.

Questions and challenges remain for future research. First, while the model predicts visual inputs in raw pixel space, it is perhaps unlikely that the navigation systems of the brain are optimized to operate at such a fine-grained level, as some details in a visual scene may not be important. Learning objectives based on higher-level visual features, for example, reconstruction of the landmarks and objects in the environment, may be more ethologically relevant. Second, the results are influenced by the agent's movement statistics. Intuitively, if the movement trajectories are highly variable, it may be difficult to predict the state for the next frame. Indeed, preliminary results reported in the paper show that increased variability of movement trajectories degrades the quality of the map learned in the model. Third, in the current model, it would be challenging to learn spatial maps under situations with weak or deprived visual inputs, such as darkness. Finally, different species show distinct behavioural strategies to sample the environment with movement and head turns. These differences may underlie variations in cognitive maps observed between rodents and non-human primates[14]. Training models such as this one with realistic movement trajectories and visual inputs specific to different species could reveal insights into cognitive map similarities and differences. A deeper understanding of the computational principles underlying cognitive map construction in the brain may enhance future embodied artificial intelligence systems.

Margaret C. von Ebers[1,2,4,6] **& Xue-Xin Wei**[2,3,4,5,6] ✉

[1]Department of Computer Science, The University of Texas at Austin, Austin, TX, USA. [2]Department of Neuroscience, The University of Texas at Austin, Austin, TX, USA. [3]Department of Psychology, The University of Texas at Austin, Austin, TX, USA. [4]Center for Perceptual Systems, The University of Texas at Austin, Austin, TX, USA. [5]Center for Learning and Memory, The University of Texas at Austin, Austin, TX, USA. [6]Center for Theoretical and Computational Neuroscience, The University of Texas at Austin, Austin, TX, USA.

✉e-mail: weixx@utexas.edu

Published online: 8 August 2024

References

1. O'Keefe, J. & Dostrovsky, J. *Brain Res.* **34**, 171–175 (1971).
2. Hafting, T., Fyhn, M., Molden, S., Moser, M.-B. & Moser, E. I. *Nature* **436**, 801–806 (2005).
3. Tolman, E. C. *Psychol. Rev.* **55**, 189–208 (1948).
4. Gornet, J. & Thomson, M. *Nat. Mach. Intell.* **6**, 820–833 (2024).
5. McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I. & Moser, M.-B. *Nat. Rev. Neurosci.* **7**, 663–678 (2006).
6. Zhang, K. *J. Neurosci.* **16**, 2112–2126 (1996).
7. Samsonovich, A. & McNaughton, B. L. *J. Neurosci.* **17**, 5900–5920 (1997).
8. Burak, Y. & Fiete, I. R. *PLOS Comput. Biol.* **5**, e1000291 (2009).
9. Cueva, C. J. & Wei, X.-X. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. In *Proc. 6th International Conference on Learning Representations* (ICLR, 2018).
10. Banino, A. et al. *Nature* **557**, 429–433 (2018).
11. Whittington, J. C. R. et al. *Cell* **183**, 1249–1263 (2020).
12. Srinivasan, M. V., Laughlin, S. B. & Dubs, A. *Proc. R. Soc. B.* **216**, 427–459 (1982).
13. Rao, R. P. N. & Ballard, D. H. *Nat. Neurosci.* **2**, 79–87 (1999).
14. Piza, D. B. et al. *Nat. Commun.* **15**, 4053 (2024).

Competing interests

The authors declare no competing interests.