No Free Lunch from Deep Learning in Neuroscience

Rylan Schaeffer [1,3] Mikail Khona [2,4] Ila Rani Fiete [3,4]

[1]Stanford Computer Science [2]MIT Physics [3]MIT Brain and Cognitive Sciences [4]MIT McGovern Institute

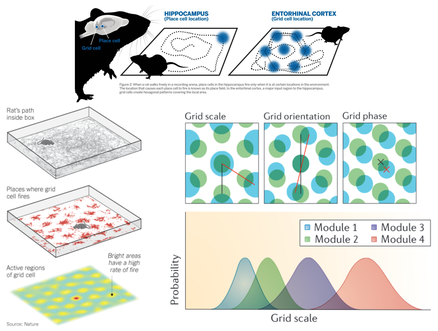
Summary

In disagreement with the main messages of [1, 2, 6, 8, 9], we demonstrate:

1. Recurrent networks trained to path integrate almost always solve the task, but almost never learn grid-like representations...
2. ...unless inserted via an intentionally-chosen & fine-tuned supervised target readouts designed to produce the grid result, meaning obtaining grid cells is post-hoc
3. The resulting grid-like units lack key properties of biological grid cells
4. The assumptions and target readouts may not be biologically plausible
5. Prior theory is suggestive, not exact or comprehensive

Prior approaches overstate the task of path integration and understate the role of the readout

How does the mammalian brain represent space?



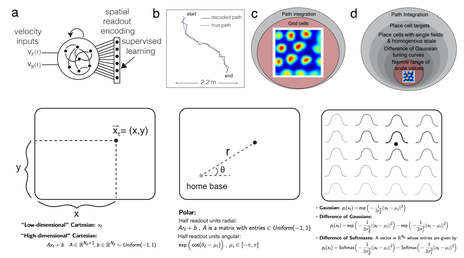
Figures from New York Times & Moser et al. 2014.

What insight(s) has deep learning claimed to offer?

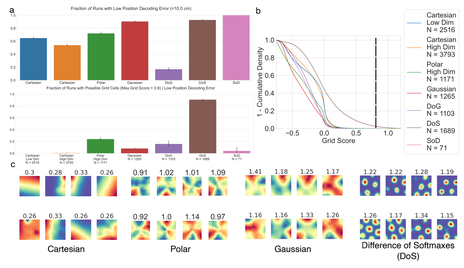
Claim: The task of path integration generically causes the formation of grid cells

- [2]: "We first trained a recurrent network to perform path integration, leading to the emergence of representations resembling grid cells, as well as other entorhinal cell types"
- [2]: "Notably, therefore, our results show that grid-like representations reminiscent of those found in the mammalian entorhinal cortex emerge in a *generic* network trained to path integrate."
- [8]: "Here we forge an intimate link between the computational problem of path-integration and the existence of hexagonal grids, by demonstrating that such grids arise generically in biologically plausible neural networks trained to path integrate. Moreover, we develop a unifying theory for why hexagonal grids are so ubiquitous in path-integrator circuits."
- [1]: "We trained recurrent neural networks (RNNs) to perform navigation tasks in 2D arenas based on velocity inputs. Surprisingly, we find that grid-like spatial response patterns emerge in trained networks."
- [9]: "RNNs trained to path integrate with nonnegative firing develop hexagonal grid cells."

Experimental Setup & Summarized Results



The task of path integration is *not* sufficient to produce grid-like representations



Grid-like units emerge only with fine-tuning one specific spatial readout

Prior theory: a theory of readout correlations

Starting with a linear readout reconstruction error,

$$\mathcal{E}(G, W) = ||P - \hat{P}||_F^2 \text{ where } \hat{P} = GW$$

[6, 9] simplifies to a Lagrangian: $\mathcal{L} = Tr[G^T \Sigma G - \lambda(G^T G - I)]$. Here, $\Sigma_{x,x'} = \sum p_i(x) p_i(x')$ is readout correlation matrix. Optimal solution to g is top eigenvectors of Σ, which for isotropically distributed, single scale readouts with DoG shape is comprised of periodic patterns. Thus:

Studying learnt representations in these networks

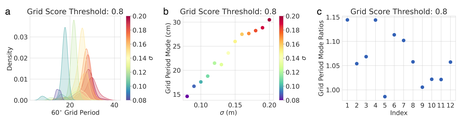
$\Downarrow$

Studying the correlational structure of readouts

$\Downarrow$

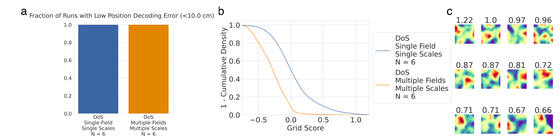
Representations learnt by these networks are wholly determined by **choice of spatial readout** (a design choice), **not task of path integration**.

Grid-like units lack key properties of real grid cells

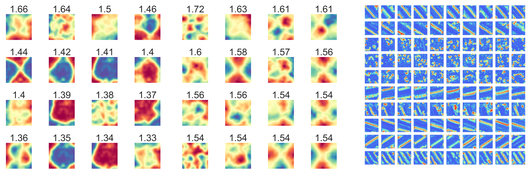


Grid periods are set by a hyperparameter & multiple modules do not emerge

With multiple scales and fields, no grid units form

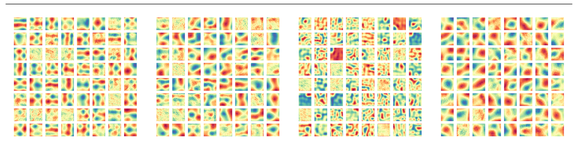


Gaussian readouts do not generically yield lattices



A: $\sigma_E = 5$ cm, sweeping arch. & seed (>12k units). B: RNNs, sweeping $\sigma_E \in [5, 50]$ cm (>177k units) by 2.5 cm incr. C: [4] independently confirm.

Gaussian readouts yield grid-like units under post hoc implementation details not captured by theory



(A) Claimed lattices from [6, 9]. (B) Rerunning degrades lattices. (C) Changing batch size or (D) removing dropout lowers loss and removes lattices.

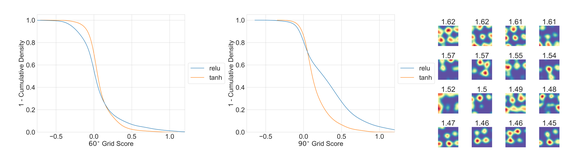
Ideal DoG place cells with non-negative activations drive hexagonal *and square* lattices

[6, 9] capture effect of non-negative nonlinearities by adding cubic term in the position-encoding Lagrangian that penalizes non-negativity ($\sigma_0 g^3$).

$$\mathcal{L} = g^T \Sigma g + \lambda(1 - g^T g) + \sigma_0 g^3$$

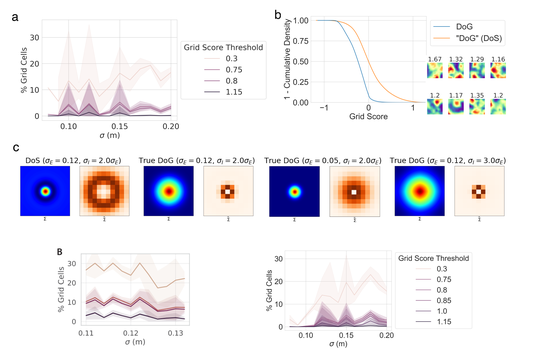
In Fourier space this term favors hexagonal patterns and thus predicts that ReLU activations favor hexagonal patterns [6, 9]:

$$\mathcal{L}_{int} = \int_{\vec{k},\vec{k}',\vec{k}''} \tilde{g}(k)\tilde{g}(k')\tilde{g}(k'')\delta(\vec{k} + \vec{k}' + \vec{k}'')$$



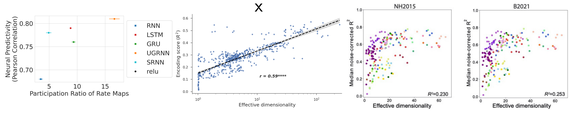
Results from recurrent network training are partially mismatched to theory

High sensitivity to implementation details



[7] contested the finding (left); rerunning with 1024 $\Rightarrow$ 4096 units reproduces it (right). This exchange fortifies claim that grids are post-hoc: all 3 sweeps learn to path integrate, thus per theory of [6, 9], all should learn grids, but post-hoc implementation details are necessary.

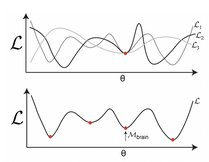
Paradox: How networks in [3] predict neural activity?



3 independent labs studying different modalities, brain regions, species, and recording methods found networks with higher participation ratio achieve higher test R^2. A: ours; B: Elmoznino et al. 2022; C: Tuckute et al. 2022.

Conclusion

Challenges with using deep learning models in neuroscience: Top: Building a model that replicates neural responses does not guarantee that optimization problem is the brain's problem, since multiple problems can share a solution. Bottom: Training networks on the plausible/correct optimization problem need not yield the brain's solution, without sufficient amounts of inductive biases.



References

[1] Cueva and Wei. *ICLR*, 2018.
[2] Banino et al. *Nature*, 2018.
[3] Nayebi et al. *NeurIPS*, 2021.
[4] Xu et al. *PMLR - Symmetry and Geometry in Neural Representations*, 2022.
[5] Schaeffer, Khona, and Fiete. *bioRxiv*, 2022.
[6] Sorscher, Mel, Ganguli, and Ocko. *NeurIPS*, 2019.
[7] Sorscher, Mel, Nayebi, Giocomo, Yamins, and Ganguli. *bioRxiv*, 2022.
[8] Sorscher, Mel, Ocko, Giocomo, and Ganguli. *bioRxiv*, 2020.
[9] Sorscher, Mel, Ocko, Giocomo, and Ganguli. *Neuron*, 2022.