

A Compressing Auto-encoder as a Developmental Model of Grid Cells

Anthony Kiggundu, Cornelius Weber, Stefan Wermter *

Introduction

The metric representation of space during navigation is attributed to grid cells in the entorhinal cortex. The cell responses form triangular grid-like patterns that tile the entire environment as an animal moves (Giocomo, Moser, and Moser [2011]). Earlier findings suggest that the precision of place cells in the hippocampus (CA1 area) of a rodent's brain is increased by the inter-connectivity from grid cells in the parahippocampal CA3 area (Moser, Moser, and Roudi [2014]). Figure 1, left, shows the grid cells organised into modules where the receptive fields of the cells in one module have the same spacing and orientation but the scale differs in others forming multiple spatially scaled modules that together precisely encode position over a large space.

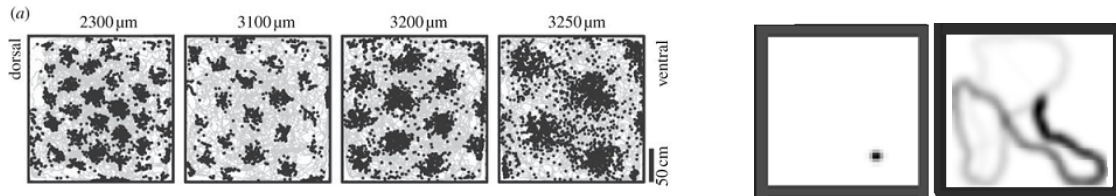


Figure. 1: Left: *In vivo* imagery of four dMEC grid cell responses in a square arena (Moser, Moser, and Roudi [2014]). The spacing of firing fields increases from left to right. Near right: Activities of place cells that serve as inputs I to our model; here, the virtual rat being at the lower-right part of the arena. Far right: a random trail taken by the virtual rat. The trail fades at the tail, the darker part showing the rat's recent path.

Although the mechanisms through which these multiple spatially scaled modules emerge are still unknown, existing neural models attribute this modular behaviour to odometry such that the change of the triangular tessellating grid cell firing is influenced by the animal's velocity and direction inputs.

In our auto-encoder model, we prescribe to evidence suggesting the existence of auto-associative networks within the entorhinal cortex which cohesively support the emerging activity patterns (Duigou, Simonnet, Teleńczuk, Fricker, and Miles [2014], Rolls [2007]). We hypothesise that grid cell responses can arise in an auto-associative model using feed-forward circuitries and inhibition mechanisms. The inhibition is implemented at both spatial and temporal level, indirectly influencing scaling and firing field sizes within the cells. The emergent grid cells carry a compressed representation of localised place cells through trained weights that encode a virtual rat's position in the environment with varying scales of grid patterns.

*Knowledge Technology Group, Department of Informatics, Faculty of Mathematics, Computer Science and Natural Sciences, University of Hamburg, Germany. weber@informatik.uni-hamburg.de

Methods

Figure 1, right, shows an example input vector to the model: simulated activities I of place cells and the trajectory of the place cells as a virtual rat randomly moves with constant velocity in a box arena. The activity is modelled as a Gaussian function centred on the position of the rat.

Our auto-encoder model is a simple feed-forward architecture with additional short-range recurrent connectivity as shown in Figure 2, left. Neurons in the input layer are connected to the hidden layer with weight matrix W^1 and the hidden layer neurons to the output layer with weight matrix W^2 . A fixed recurrent or lateral weight matrix W^3 implements short-range spatial inhibition. W^1 and W^2 are randomly initialised, bias vectors b_1 and b_2 are added to the hidden and output layers. The size of the input space is 1600 (40×40) neurons, the hidden layer has 16 neurons and the output space the same size as the input. This forms a compressing auto-encoder with strongly under-complete coding. The output layer activation is complemented with a competitive softmax function to let only those place cells fire for which grid cells of several different scales agree.

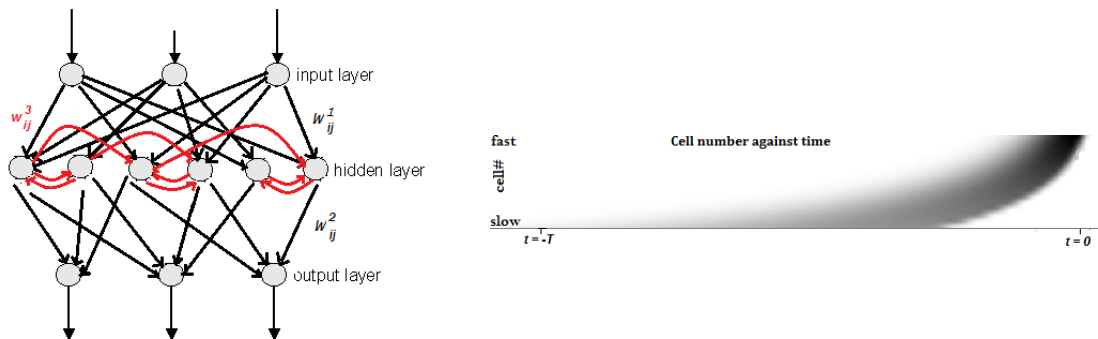


Figure 2: Left: Auto-encoder architecture. W^3 (red) is the short-range recurrent cell-to-neighbourhood connectivity for spatial inhibition. Right: The temporal inhibition function $H_{(t,i)}$. The plot shows cell number (y -axis) over time (x -axis). Grey indicates self-inhibition of an activated neuron after time t . Faster cells (top) receive activity inhibition from recent history; slower cells from activity deeper back in time.

Since we assume no prior knowledge of space, we implement a temporal inhibition mechanism, which is based on the notion that grid cells of high spatial frequency will be quickly activated and deactivated as a rat moves, while cells of low spatial frequency have slow activity changes. The inhibition mechanism allows cells to remain active only for limited times. Figure 2, right, shows the function $H_{(t,i)}$ which determines how much temporal inhibition h_i neuron i receives from its previous activations S_i , inhibiting fast cells more quickly from their own activities than slow cells:

$$h_i(t) = \sum_{t'}^T H_{it'} \cdot S(t - t') \quad (1)$$

where T is the memory span, t is current and t' previous time-steps.

The spatial inhibition via short-range inhibitory recurrent weights W^3 causes distant neurons to fire independently. The net hidden layer S activity was then computed by applying a sparse transfer function g .

$$a(t) = W^1 \cdot I(t) + W^3 \cdot S(t - 1) + b_1 - \eta \cdot h(t) \quad (2)$$

$$S = g(a) = a - 0.9/(1 + a^2) \quad (3)$$

where η scales the temporal inhibition. Activation on the output layer is computed as

$$O = \text{softmax}(W^2 \cdot S + b_2) \quad (4)$$

where W^2 are the weights to the output layer with the respective bias vector b_2 . The error on the output layer $e = I - O$ is then used for learning of the weights by back-propagation using gradient descent on the sum square error.

Results

Figure 3 shows the emergent weights of the 16 hidden layer neurons after 70000 training steps. Receptive fields of the cells are spatially organised in approximately triangular grids, showing grid cell responses. The scales of these grids increase from left to right, i.e. from grid cells with faster temporal inhibition to cells with slower temporal inhibition.

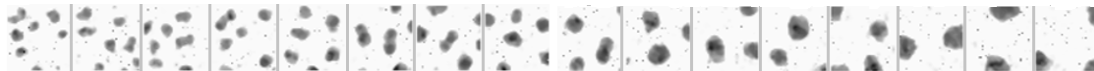


Figure 3: Model results. Each square shows the input weights of one of the 16 hidden layer neurons, i.e. one row of W^1 . Non-zero weights (dark) connect to isolated regions in input space forming triangular arranged grid patterns, which vary in size, from small scales of "fast" grid cells (left) to larger scales of "slow" cells (right).

Discussion

We implemented an auto-encoder that encodes a localised place cell input efficiently with fewer grid cells. Varying temporal local inhibition led to varying grid spacing, while spatial short-range inhibition enforced the coding to be performed by cells of different scales. The results simulate the emergent triangular grid pattern activity at different scales where the cells' receptive field weight profile (Figure 3) is similar to biological findings of grid cell activations (Figure 1, left).

Our simple model does not integrate path signals from odometry to influence the behaviour of the activity bumps, as most other models of grid cells do. Nevertheless, the emergent hexagonal grid patterns are stable over time, explaining emerging and convergent connectivity between place cells and grid cells, which existing models do not yet explain. Naturally, odometric information needs to be represented within the hippocampus/dMEC to achieve accurate path integration. Introducing odometry could lead to a class of more detailed and plausible grid cell models.

References

- C. L. Duigou, J. Simonnet, M.T. Teleńczuk, D. Fricker, and R. Miles. Recurrent synapses and circuits in the CA3 region of the hippocampus: an associative network. *Frontiers in Cellular Neuroscience*, 7:262 – 280, 2014.
- L. M. Giocomo, M. B. Moser, and E. I. Moser. Computational models of grid cells. *Neuron*, 71:589–604, 2011.
- E. I. Moser, M. B. Moser, and Y. Roudi. Network mechanisms of grid cells. *Phil. Trans. R. Soc. B*, pages 369 – 376, 2014.
- E.T. Rolls. An attractor network in the hippocampus: Theory and neurophysiology. *Learning & Memory*, 14:714 – 731, 2007.