Learning Boundaries in Space Using Convolutional Neural Networks

6.884 Introduction to Machine Learning Final Paper *

Cagri Hakan Zaman
zaman@mit.edu

Ege Ozgirin
egeozin@mit.edu

May 15, 2016

Abstract

Recent discoveries in neuroscience suggest that a variety of cell groups in the hippocampus (i.e. place cells) and entorhinal cortex (i.e. grid cells, head direction cells and boundary cells) provide a representation of space that act as a cognitive map. This representation combines allocentric sensory cues with proprioceptive information, allowing the animal to capture invariant features of the environments and robustly navigate between different places. Our project explores machine learning techniques that can represent this architecture and to develop a working model in the context of robotic perception and navigation.In this paper we introduce a biologically inspired architecture that allows for robust visual inference and navigation in novel environments using convolutional neural networks (CNN) After an online training phase with a series of distance sensors, our model predicts activation patterns for the boundary cells when presented an image. Using predicted boundary cell activations the robot is able to navigate in the simulated environment while avoiding obstacles with a very high success rate.

1 Introduction

Our brains spend a tremendous effort to resolve space, integrate the sensory information into practical information so that it can plan where to move, where to sit; remember a particular place so it can navigate back when it is necessary. Basic characteristics of spatial perception have been discovered through studies on rat hippocampus. (O'Keefe and Nadel, 1978; Ranck, 1973; Emery, Wilson and Chen, 2012). Different firing patterns in place cell region allow the animal to represent its location relative to the objects in the environment. While place cells respond to different visual cues in the environment, they also integrate proprioceptive information as to direction and velocity, as well as the sensory information other than visual ones such as sound or odor. Spike pattern of a particular place cell is correlated with the change in size, orientation and shape of the environment (O'Keefe, 1998).

*The note about the collaboration between the students can be found at the end of this paper.

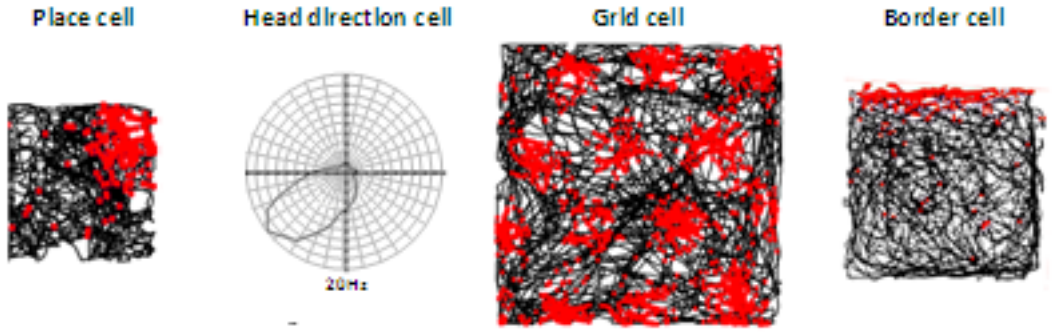


Figure 1: Spiking patterns of spatial cells of a rodent brain in a square environment. A place cell has a "place field" and is activated only in a particular part of the environment. A head direction cell is activated only when the rodent is oriented towards a particular direction in the environment. A grid cell has an activation pattern that has a particular distance interval which builds a "grid" when recorded in multiple locations. A boundary cell (border cell) is activated when the rodent is nearby a boundary towards a particular direction. In this example, the boundary cell is activated nearby the north wall of the room

In addition to evidently location selective place cells, there are other neural compositions in brain such as grid cells, which have special grid-like firing fields in the environment; head direction cells, which are activated when the rodent is facing their selected direction; and boundary cells, which is activated when the rodent is nearby a border facing a particular direction. These cells contribute to the formation of reliable information about the environment (Figure 1). Additionally, PPA area in human brain is selective to environmental structures, which integrates low level visual features to higher level descriptions that allow detection and recognition of individual spaces.

Visual space learning poses an important challange for developing artificial systems. In a previous study, Oliva et.al showed that perception of a scene rely on several components including low-level visual features,previous experience about the environment, and spatio-temporal history in the environment (Oliva et.al, 2011). According to their study, perceptual information regarding a scene can be characterized as the spatial-envelope, an object level description of scene geometry, such as openness, roughness, perspective and volume (Oliva and Torralba, 2001, 2006). Integrating these features into a place representation requires bridging multiple descriptions obtained from an active exploration of space. We are also inspired by a few successful implementations of simultaneous location and mapping (SLAM) methods, notably the RatSLAM (Milford et.al, 2010), which uses a place-cell inspired model for generating maps in novel environments. RatSLAM uses place cell representations to assign camera images to unique places where the linear motion and orientation information is extracted from images through an optical flow method.

In this project we use a deep learning architecture to extract boundary cell activations from camera images. The main difference of this approach from other SLAM models is that it does not generate holistic maps, topological representations or geometric models. Instead, in any given time, we predict boundary cell activations using only the provided camera image. This method allows efficient decision making for robot navigation as CNN tends to produce predictions very robustly. We test our system for two tasks. First, we use boundary cell predictions to decide the

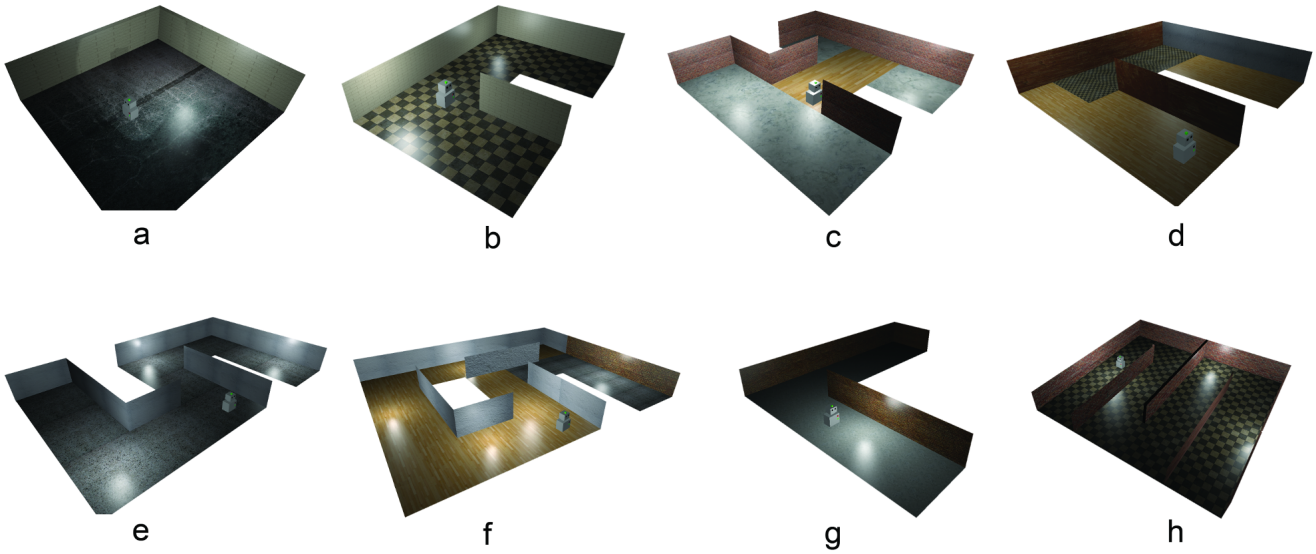


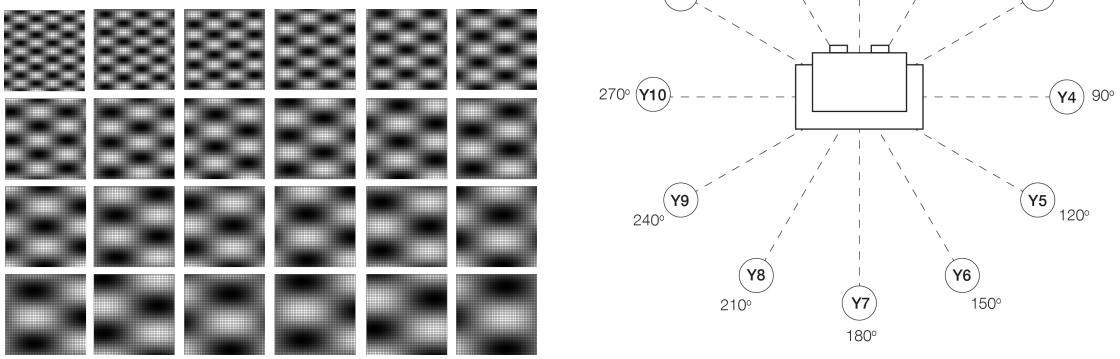
Figure 2: Our model was trained and tested in eight different simulation environments. Five environments were selected for training while three of them were used for testing.The test environments included novel textures that were not included in training environments.

most plausible path for the robot to take in the next frame so that it avoids obstacles. Results from three testing environment show that our method allows robot to successfully avoid obstacles and randomly explore the environment without supervision. Second, we use boundary activations to determine unique locations in the environment. We show that the predicted activations of our model are selective to obstacles in particular allocentric orientations,(e.g.North faced walls), which presents similar characteristics to the recordings from boundary cells in rat brain.

In the scope of this paper, we only show the discriminative characteristics of the boundary activations, however, research suggest that place cell activations benefit from integration of boundary, grid and head direction signals.As a first step for the future implementations, we trained another model for predicting grid-cell activation. However, this model did not produce reliable predictions. We think it is intuitive, because grid cells use self movement as the main input where allocentric cues —such as the camera image that we used in our model— have a minimal impact on grid cell activations. We believe a secondary system that rely on movement of the robot can be used for generating grid activations and these two systems (boundary and grid) can be integrated to predict place cell activations. At the final section of our paper we outline such an implementation for place recognition.

2 Methodology

As a first step, we have created a robot simulation environment with the MORSE library in Python/Blender. We designed five environments for training and three environments for testing (Figure 2). Our simulated robot is supplied with an onboard camera and a laser scanner. We wrote several utility functions to form appropriate representations that are isomorphic to the spatial layers -boundary cells, head direction cells and grid cells in rodents. We formed four different models by implementing a CNN architecture with four different types of label vectors. First three types are



(a) Grid Cell Features

(b) Boundary and Head Direction Features

Figure 3: Feature Representations. a) Grid maps are used to train the system to predict grid activations. There are 24 different grids that have different phase and frequency. Final grid feature G was calculated with a function $F(G_t, x, y)$ where $t \in [1..d]$ and x and y are real valued positions obtained from simulation.b) Head Direction and Boundary features are selective to orientations. Activations of these cells calculated with functions $F(B_t, \alpha)$ and $F(H_t, \alpha)$ where $t \in [0...d]$ and $\alpha = t * 2\pi \div d$

8, 12, 24 dimensional vectors respectively, representing multiple boundary cell activations. Fourth label type is a 24-dimensional vector representing 24 grid cell activations. We trained our model separately with boundary cells and grid cells.

2.1 Spatial Cell Representations

2.1.1 Boundary Cells

We used the output values of the SICKLMS500 laser scanner model to create boundary cell activations. The data from the sensor is provided in an array that stores the distance to the first obstacle received from each ray. Boundary cell activation values are generated by mapping the laser-scanner output between 1 and 0. The values are close to 1 if the agent is in close proximity to a wall and close to 0 if it's distant from a wall. The value of a boundary feature B is given by:

$$F(B_t, \alpha) = \begin{cases} 1 & \text{for } L(\alpha) \leq \sqrt{k} \\ k/L(\alpha)^2 & \text{otherwise} \end{cases}$$

where $L(\alpha)$ is a function that gives a distance reading from the laser scanner from the direction α and k is a constant value for the minimum distance for a cell to reach the maximum activation.

2.1.2 Head Direction Cells

Agent's pose could be gathered by using a pose sensor providing the rotation values around the axes of the sensor. Head direction cell activations are initially mapped from the radian values

received from the sensor to the cartesian values. We calculated a gradient for each of the 12 cells by representing the most active cell with a value closer to 1 and assigning the less active cells with lower values that are closer to 0 relative to the most active cell.

2.1.3 Grid Cells

Finally, 24 grid cell activations are generated by computing a gradient of frequency and phase values that are consistent with biological data. 24 different grid cells are triggered only when agent transpasses grid activation zones in the environment based on each cell's grid spacing and phase shift parameters that vary between 10 cms to 100 cms. Again, the output values of grid cells are designed to be real values between 0 to 1, where 0 means the cell is dormant and 1 means the cell spikes in the maximum rate. Activation of grid features are calculated by:

$$G_t(F, P, x, y) = 0.5 * \sin(P_t + 2 * \pi * x/F_t) * \sin(P_t + 2 * \pi * y/F_t)$$

where phase delay $P_t \in [0..\pi]$ and grid frequency $F_t \in [0.1..1]$ (Metric units between 0.1m to 1m).

Making the location available for grid cells does not contradict with the idea of inferring place information, as current biological evidence supports that grid cells are innervated with neurons carrying proprioceptive information. The question of how the location information becomes available for the grid cells is particularly interesting but it's outside of the scope of this study.

2.2 Model

We developed a learning model for the agent to identify unique places in a given environment. Our aim was to make the agent rely only on the visual input to 1) determine its relative location in the environment and 2) navigate to goal locations. Our system employs a custom CNN architecture in which the output layer produces a vector Vs that represents the activation of spatial layers - boundary cells, head direction cells and grid cells. Our CNN architecture consists of two consecutive convolution and max pooling layers. Convolution layers use 32 3 x 3 filters and ReLU activations. Max pooling layers have the size 2 x 2. Finally we apply two fully connected layers before the output layer. Output layer dimension varies according to the training variation. We test $d = 8, d = 12, d = 24$ for the dimension of feature vector Y in three different training sessions. In all sessions, we used sigmoid activation for the output layers and root mean squared objective function.

All labels were real valued numbers such that $Y_i \in [0, 1]$ each corresponding to the level of activation of a particular cell. Each feature vector Y was independently generated and does not correspond to a predetermined class. Therefore, unlike CNN's that perform classification tasks, probabilistic representations such as $softmax$ was not suitable for our purposes. As our feature vector Y is a random vector that is continuously generated from the environment during the online training phase, we decided to use mean squared error for the loss function. Additionally, we experimented with different CNN architectures and learning parameter values.

The model is trained online with camera images that are streamed from the simulation environment with a rate of 2 frames per second. Images are converted to $256x256x3$ RGB format and normalized to floating point values such that $X \in [0, 1]$. In each frame, label values are calculated from the sensors, and and CNN is updated with this labels and given the camera image. We perform two turns of back-propagation in each update.

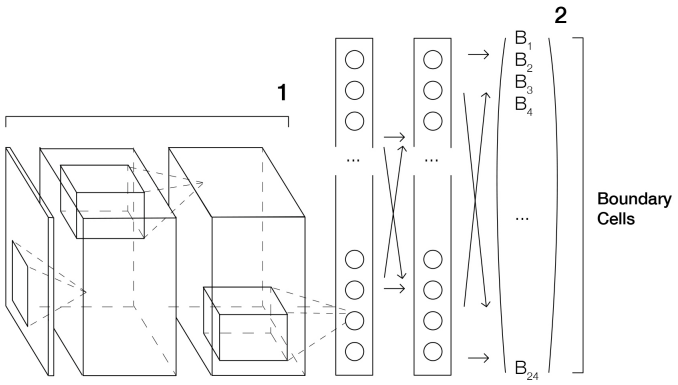


Figure 4: CNN Architecture for predicting Boundary Cell Activations. This model produces a vector of values in range [0,1] that correspond to the level of activity for an individual boundary cell for a particular direction.

2.3 Online Training

We trained the system in 5 different environments with the label vectors composed of 8, 12, 24 boundary cells and 24 grid cells respectively in four different settings. Training phase consisted of three-minutes training of the model in each simulation environment for three cycles. In the first two cycle, the model was trained while the agent navigates in the environment using distance data provided by the laser-scanner. In the third cycle, the model was continued to be trained while the agent navigates based on the prediction of the model. In the last cycle, researchers interfered with the process whenever the robot failed to detect a boundary and hit the wall.

The training procedure is performed consecutively in each environment in each cycle. We changed the environment every three minutes and continue training the network. Because there was no way of presenting randomly selected samples to the network as in an ideal scenario, we believe altering the environments between cycles allowed us to prevent over-fitting to a particular setting and provided a better generalization. Figure 5 shows the loss value after each sample over the all training sessions. We observed the loss value tends to get closer to a value 0.02 while fluctuating up to 0.08.

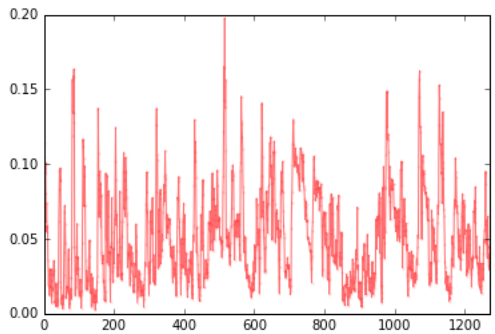


Figure 5: Loss value after each sample over the all training sessions. The final model has fluctuating loss values between 0.02 and 0.05.

3 Results

3.1 Boundary Prediction

We trained three different networks using only boundary features of size $d = 8, d = 12, d = 24$. Each network was tested in three novel environments that were not included in the training sessions.The network trained with $d = 24$ failed to avoid obstacles and presented a very poor performance for predicting boundary activation. Because it was not possible to record the predictions without robot avoiding the obstacles, we excluded this model from our analysis. Networks with $d = 8$ and $d = 12$ successfully navigated in the test environments. Please refer to this video for a sample test session.[1]

In order to compare the performance of our models predictions, which are relative to robot's orientations, with biological boundary cells, which are tied to global orientations, we have plotted predictions regarding four major directions. Figure 6 shows a comparison of robot trails in two models, d=8 and d=12 for North, South, East and West selective cell activations predicted by the system. Part A, predictions from 8-cell model, shows a better discrimination for boundaries compared to Part B, the 12-cell model. However, in both of the models, there are many parts where the cell shows high activation although there is no boundary in the selected direction.

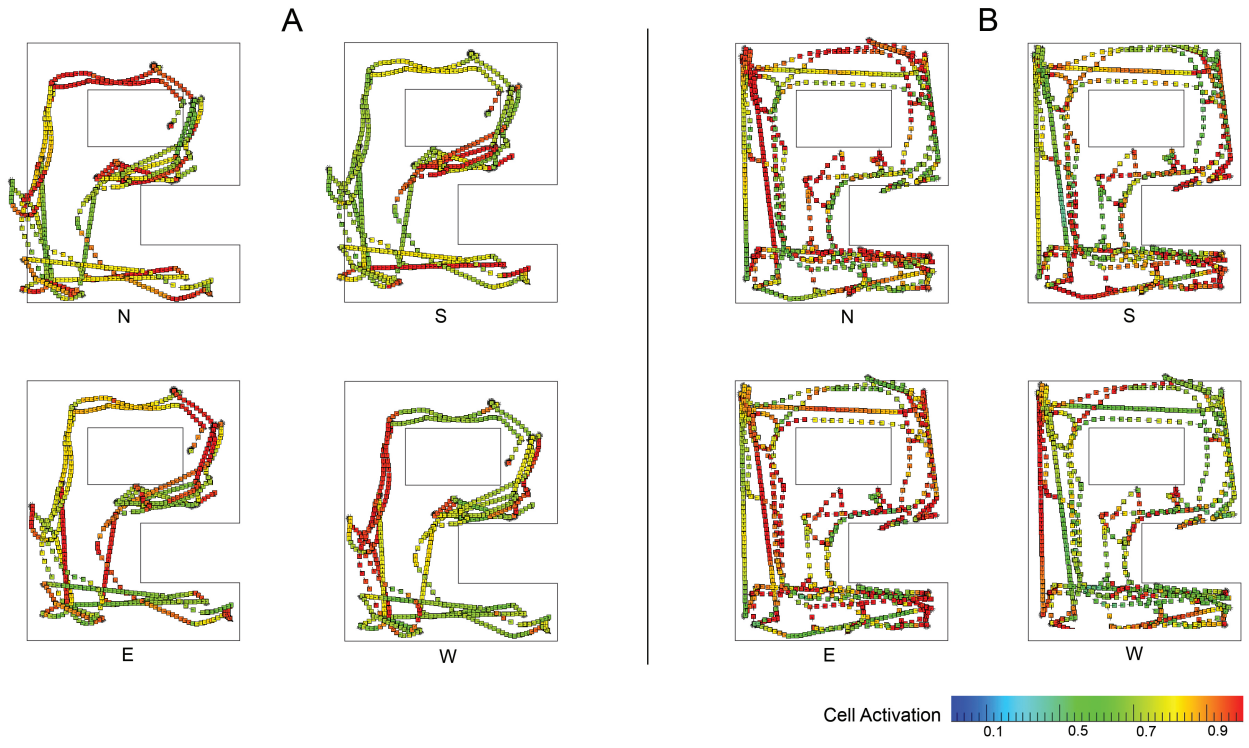


Figure 6: A. 8-Cell Model predictions for four different directions. It can be seen activations are higher around the borders for the indicated directions. B. 12-Cell Model predictions. This model shows less accuracy for predicting activation values.

[1]The video file can be found at: https://vimeo.com/166680547

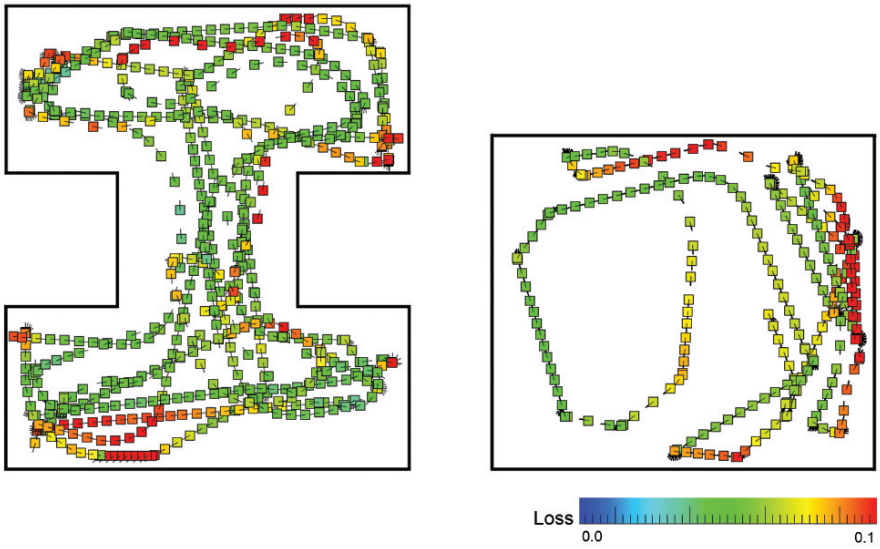


Figure 7: Loss values for the predicted activation values in two different training environment.It can be seen that the loss value gets higher nearby the borders.

Figure 7 shows the loss value of predictions for two different training environments. We observe that the loss value gets higher around the borders. We believe there are two possible explanations for this behaviour. Because the feature values are generally smaller in central areas, the error in prediction would be getting relatively smaller as well. Around the borders there are more cells with maximal value which would produce a higher L2 distance between actual and predicted values. Secondly, the camera image changes rapidly nearby the borders because of the distance and gets more homogeneous especially if the robot is faced towards the wall. The system would be more prone to errors nearby borders. We believe a better distance rating would be used for addressing this issue.

3.2 Grid Predictions

Finally, we have tested a model we trained with 24 grid features to test if we can predict grid cell activations with the same model. Figure 8 shows three different grid cell predictions from the model that is trained for 15 minutes in a square environment. As it can be seen on the figure, the

Figure 8: Predicted activations for 3 different grid feature in a square environment. Refer to Figure 3 for corresponding input Grid Maps. The predictions are not reflecting the expected patterns although they present some grid-like regularity.

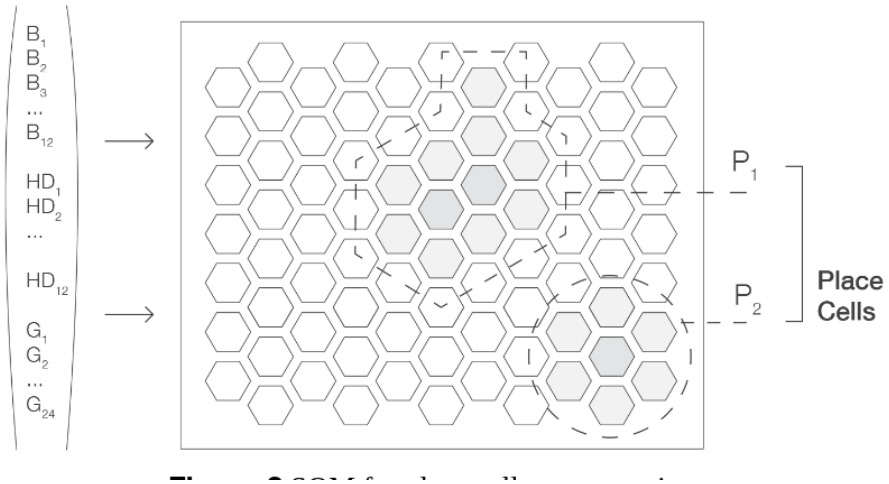


Figure 9: A Self Organizing Map (SOM) can be used to generate unique place descriptions from boundary, grid and head direction cells.

predictions did not converge to required grid patterns, although there are partial repetitions in the spike pattern. We found it intuitive as grid pattern is orientation invariant and mostly depend on proprioceptive signals rather than visual ones. However, this model can be complementary to a prediction system for calibration and error correction.

4 Discussion

4.1 Future Work

We plan to implement a self organizing map (SOM) that represents place cell activations for identifying unique places in the environment. SOM is used for navigating from one place to another without using a global reference (i.e. a cognitive map). We also think that a simple HMM can be employed for predicting the sequence of places. We will compare the results obtained with these two different techniques and direct our future studies accordingly.

4.2 Contributions

Although several biologically-inspired learning approaches exist for modeling spatial navigation and representation(Milford et al. 2009), our model differs in terms of learning strategy and architecture. Initially, our model employs a unique learning approach by training itself online with dynamic labels representing cell activation values that are sampled from a particular random distribution. Moreover, we demonstrated that a CNN with two convolutional layers provides a decent mechanism for predicting cell activation values from images.

In conclusion, through this project we have provided a model to understand mammalian spatial representation and inference mechanisms contributing to the field of computational neuroscience and cognitive science studies. Furthermore, we implemented a multi-modal visuospatial learning framework that could be useful for scene recognition and robotic navigation.

References

[1] Park, S.Brady, T.F.Greene, M.R., Oliva, A. 2011.*Disentangling scene content from its spatial boundary: Complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes* Journal of Neuroscience, 31(4), 1333-1340.

[2] Oliva, A., Torralba, A. 2001. *Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope.* International Journal in Computer Vision, 42, 145-175

[3] Oliva, A. Torralba, A.2006.*Building the Gist of a Scene: The Role of Global Image Features in Recognition.* Progress in Brain Research: Visual perception,

[4] O'Keefe, J. and Nadel, L. 1978. *The Hippocampus as a Cognitive Map. Oxford: Oxford University.*

[5] O'Keefe J Burgess N Donnett J G Jeffery K J Maguire E A .1998 *Place cells, navigational accuracy, and the human hippocampus* Philos Trans R Soc Lond B Biol Sci. 1998 Aug 29; 353(1373): 1333–1340.

[6] Ranck JB., Jr. 1973.*Studies on single neurons in dorsal hippocampal formation and septum in unrestrained rats.* I. Behavioral correlates and firing repertoires. Exp Neurol.41:461–531.

[7] Chen, Z, Kloosterman, F, Brown E N, Wilson, M A,2012. *Uncovering spatial topology represented by rat hippocampal population neuronal codes* Journal of Computational Neuroscience, October 2012, Volume 33, Issue 2, pp 227-255

[8] Milford, M.,2010. *Persistent Navigation and Mapping using a Biologically Inspired SLAM System* The International Journal of Robotics Research vol. 29 no., 1131-1153

[9] Chen, Z, Kloosterman, F, Brown E N, Wilson, M A,2012. *Uncovering spatial topology represented by rat hippocampal population neuronal codes* Journal of Computational Neuroscience, October 2012, Volume 33, Issue 2, pp 227-255

5 Note on Collaboration

The selection of the problem, and background research was carried by both of us.

Ege has designed and implemented the CNN architecture. He designed the simulation environments. He developed procedures for handling image data and sending it to CNN model. He produced architecture diagrams and loss plots.

Cagri has implemented the sensor data stream and feature mapping from the simulation environment. He has implemented the robot navigation procedures for test and training sessions. He produced cell activation plots.