Grid Cell Path Integration For Movement-Based Visual Object Recognition

Niels Leadholm[1,2]
niels.leadholm@seh.ox.ac.uk
Marcus Lewis[1]
mlewis@numenta.com
Subutai Ahmad[1]
sahmad@numenta.com

[1] Numenta, Inc.
Redwood City,
California, USA

[2] Dept. of Experimental Psychology
University of Oxford
Oxford, UK

Abstract

Grid cells enable the brain to model the physical space of the world and navigate effectively via path integration, updating self-position using information from self-movement. Recent proposals suggest that the brain might use similar mechanisms to understand the structure of objects in diverse sensory modalities, including vision. In machine vision, object recognition given a sequence of sensory samples of an image, such as saccades, is a challenging problem when the sequence does not follow a consistent, fixed pattern - yet this is something humans do naturally and effortlessly. We explore how grid cell-based path integration in a cortical network can support reliable recognition of objects given an arbitrary sequence of inputs. Our network (GridCellNet) uses grid cell computations to integrate visual information and make predictions based on movements. We use local Hebbian plasticity rules to learn rapidly from a handful of examples (few-shot learning), and consider the task of recognizing MNIST digits given a sequence of image feature patches. Extending beyond the current literature, we show that GridCellNet can reliably perform classification, generalizing to both unseen examples and completely novel sequence trajectories. Furthermore, by utilizing grid cells for an internal reference frame derived from sensory inputs and internal motor information alone, the classification process represents an important step towards enabling translation invariance in sequential classifiers. In addition, we demonstrate that GridCellNet is able to predict unsensed regions of the input, that inference can be successful after sampling a fraction of the input space, and that a natural benefit of the proposed architecture is robustness in the context of continual learning. We propose that agents with active sensors can use grid cell representations not only for navigation, but also for robust and efficient visual understanding.

1 Introduction

When exploring a visual scene, primates sample the world in a serial sequence by performing rapid eye movements known as saccades [27]. For the purpose of recognising objects, it is non-trivial that this sampling can follow an arbitrary sequence order, and begin on any part of the object. For example, one might selectively attend to the most salient parts of a face rather than performing a raster scan across the image. While many previous efforts to model primate object recognition have focused on massively parallel processing of a single input, the challenge of dealing with the necessarily sequential nature of sensory inputs has received

less attention [3]. Recurrent-neural networks can perform complex tasks with sequential inputs, and might seem like a natural candidate for such a challenge, yet they struggle to learn when provided with sequences that do not follow a fixed order during training and inference. In the natural world there are additional challenges that can present themselves. Often only a handful of object examples are available, learning should be rapid (i.e. requiring limited training on the few examples given), and representations should be robust in the face of future learning. These are all constraints that humans are able to handle effortlessly. Understanding how learning and inference under these conditions might be achieved has two appealing aspects. As well as potentially uncovering the basis for human performance in this domain, the flexibility to operate under such a regime could also enable artificial agents to explore the world in a more principled and adaptive manner.

While recurrent neural networks do not have explicit mechanisms for dealing with this challenge, grid cells might provide a neurally plausible solution. Together with place cells in the hippocampus [20], grid cells in the entorhinal cortex enable the brain to model space during navigation. In particular, grid cells fire in repeating patterns as space is traversed [6]. Using multiple grid cells of different scale and orientation, the location of an animal can be uniquely encoded [5]. Importantly, this location representation can be updated to support path integration - that is, given information about self-movement, an agent can determine its new location by reading out from grid cell activity [6, 17]. The role of such cells in spatial navigation is widely established, but recent experimental evidence has also uncovered the presence of grid cell-like activity in visual space [11, 12, 19]. Theoreticians have argued that grid cell-like computations might be used to build object representations in diverse sensory modalities [8], including vision [3]. This is an intriguing solution to our opening problem, but the demonstration of object recognition with such computations has so far been limited to either synthetic objects [15], or visual tasks requiring the recall of a memorized example [3], rather than generalization to unseen examples of an object class.

Neurally-motivated systems that can solve rapid object learning and recognition given saccade-like visual inputs are therefore lacking. We set out to address this by implementing a biologically plausible network, called GridCellNet, based on cortical columns and grid cell-like computations. The system uses rapid Hebbian-style learning to associate sensed features and their spatial location in the reference frame of an object, while dendritic segments enable the system to encode predictive states. Locations are encoded by activity in grid cell modules that are updated with self-movement.

GridCellNet addresses the challenge of arbitrary sequence inputs, and due to the use of an internal reference frame for representing space, marks an important step towards classification that is translation invariant. Furthermore, the system includes predictive capabilities, enabling completion of an image given partial inputs. Finally, GridCellNet's properties naturally confer rapid learning (functioning with both few training examples and few weight updates), and robustness to learning additional classes, while retaining classification performance on older classes (continual learning). We evaluate the performance of GridCellNet in these task settings, and compare it to typical machine-learning approaches. While our evaluation is limited to MNIST [14], robustness on this simplest of visual data-sets is far from resolved [18, 23, 25]. In accordance with human capabilities, our system outperforms more traditional machine learning approaches in the challenging setting we explore.

To summarise, our primary contributions are to:

- Implement a biologically-motivated architecture that uses arbitrary sequences of local visual features across space to learn objects and recognize them. We provide evidence

that this ability is dependent on integrating self-movement for sensory predictions.

- Demonstrate the ability of our network to successfully generalize to unseen objects given a series of sensations with arbitrary starting positions and sequence order, a necessary element for translation invariance, and a form of out-of-distribution generalization. We show that other machine learning systems, such as recurrent neural networks, struggle by comparison on limited training examples (few-shot learning).
- Demonstrate additional benefits of the proposed architecture, such as robustness under continual learning, and the ability to predict unsensed regions of the input.

1.1 Related Literature

As noted, the demonstration of object recognition using grid-cells has so far been limited to either synthetic stimuli [15], or the recall of memorized examples [3], rather than generalization to novel samples of an object. Related work in robotics that uses sensory and self-movement information has also been limited to the recall of training examples [9, 21]. This is where our work most significantly builds on the prior state-of-the-art.

We also contrast our approach to previous methods in multi-view object recognition in machine learning. These systems often assume that there is no spatial structure to the disparate sensory inputs, instead aggregating them in an agnostic manner [22, 24]. Alternative approaches that do make use of spatial information do so using an external reference frame (such as coordinates in the input space) [13, 16, 26], which constrains their ability to generalize to out-of-distribution locations. By using an internal reference frame to perform inference, our approach makes use of spatial information while avoiding the limitations of relying on a fixed, external reference frame. An extended discussion of work related to our approach is provided in the Supplementary Material (Related Literature).

2 Methods

2.1 Overview

Our work builds on the sensorimotor system implemented in Lewis et al. [15], which in turn uses many of the algorithms employed in Hawkins et al. [7]. In this paper we address two limitations of the network in Lewis et al. [15]. First, they used synthetic objects and features rather than those derived from natural data-sets. Second, their system was designed to only recall previously seen objects and did not generalize to other examples of an object class. In order to remove those limitations, we make two primary changes:

- We implement a sparse convolutional feature detector, trained on images, where the extracted features approximate a series of foveations across an image. When the sensorimotor system moves to sample an image patch, the corresponding subset of sparse feature outputs is sent from our convolution based pre-processing.
- We modify the main classification step, using multiple learned location representations as features for object recognition. Classification then operates on the set of active grid cells using a linear transformation followed by a normalization process.

Except where noted, the sensorimotor network is mathematically as described in Lewis et al. [15], and we advise readers interested in those details to refer to that work. The details of the alternative models we evaluate are provided in the Supplementary Material (Methods).

2.2 Sparse Feature Extraction

In order to handle realistic images, we use a trained convolutional neural network to generate sparse binary features at multiple image locations. Intuitively, this approach is intended to create a map approximating how a fovea might represent the input at different regions. Note this pre-processed input to our downstream classifiers is therefore not treated as a retinotopic map. A k-Winner Take All (k-WTA) layer [1] is used to enforce sparsity in the middle-layers of this encoder network, and these representations are binarized before being passed to the downstream classifiers (i.e. both GridCellNet and the classifiers we compare to). The encoder network is described in more detail in the Supplementary Material (Methods).

2.3 Sensorimotor Network

The sensorimotor network consists of two layers, one representing the sensory input, the other the location of the sensor. Cells in both layers can be either on or off, and activity in the network proceeds through a series of discrete time steps. The sensory layer receives the input features, as well as modulatory input from the location layer. The location layer receives movement information, as well as input from the sensory layer. The connections between the sensory and location layers are modelled via dendritic segments, the small branches on biological neurons that integrate multiple synapses [2]. A dendritic segment is deemed active if there is a significant match between the sending layer's sparse activity and sparse, learned weights. This match must exceed a user-set threshold, after which these dendritic segments enable a given layer's activity to predict representations in the other layer. For classification in our task, we use the term 'object' when referring to a particular instance of a hand-written digit, and 'class' when referring to the identity of the digit.

The basic intuition for inference is that at the outset, a sensory feature is likely to be ambiguous as to the nature of the object, and so the network should encode this ambiguity with a representation that corresponds to a union of all the objects compatible with this given feature. For example, a curved contour at the top of an image might represent a 9 or a 0 (Figure 1a). The object-representation is encoded by the activity in the location layer (as each learned object uses a unique location space), and so this union of multiple objects will correspond to multiple cells being active in each grid cell module. In our example, the active location representation will correspond to both where a curved contour was learned for a 9 and where it was learned for a 0 (Figure 1b). As additional features are sensed, the network will use its current representation of candidate objects to predict the next feature, with only those that are compatible with the subsequent sensation persisting. Notably, these predictions rely on the presence of a feature *at a given location*, and not simply a bag-of-features detector. If the sensor was to move to the bottom of the image, the learned 9 representation and 0 representation will predict different features. As GridCellNet experiences more sensations, its location representation should converge to a more limited number of learned objects. The object is categorized when the representation in the location layer corresponds predominantly to a particular class, and not any others. This class-correspondence is determined by a learned linear transformation of the active grid cells, described in the following section. We provide additional details on the sensorimotor network architecture and its implementation of Hebbian learning in the Supplementary Material (Methods).

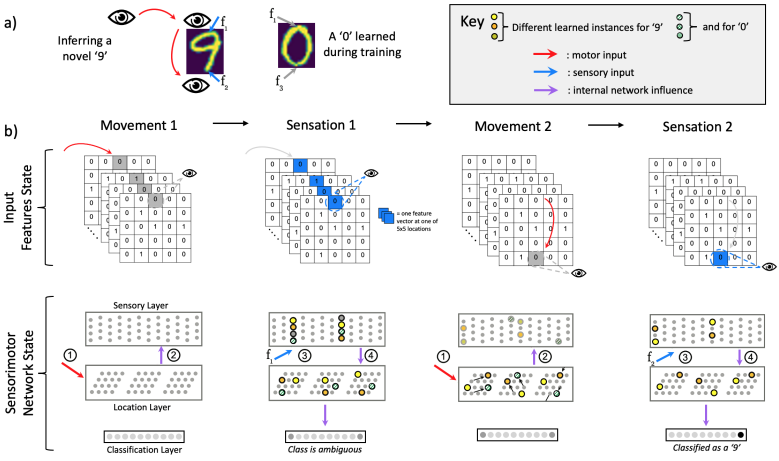

Figure 1: *Overview of GridCellNet's classification algorithm.* a) The algorithm's concept at a high level. A single sensation (f_1) is likely to be ambiguous as to the nature of the object. Correct inference requires integration over several features, such as a movement to the location of f_2. b) Sparse feature vectors extracted from a CNN feature map (upper row) represent an approximation of 5×5 foveations of the input. These sensations and associated movements are sequentially provided to GridCellNet's sensorimotor network (bottom row). When the sensor moves to its first location (<1>, under Movement 1), there is no location representation on which to base sensory predictions (<2>). As a result, the first sensation (f_1), activates all the cells in columns that receive input (<3>). This activity then activates location representations which are consistent with that feature input (<4>). This will be a union of multiple object representations, some of which are consistent with the target class (shades of yellow), and some of which are not (textured shades of green). As multiple class nodes receive active input, with no clear winner, the class identity of the object is ambiguous. With the next sensor movement (<1>, under Movement 2), the location representations of the grid cells are updated using path integration. The active grid cells then provide a prediction (via the modulatory impact of dendritic segments) to the sensory layer (<2>). The next sensory input (<3>) is consistent with two previously learned examples of 9's, i.e. with that feature representation at that location *given the previous feature representations at their locations*. The remaining active location representations (<4>) now drive a single target class above a threshold relative to the activity of the other class nodes. Note that Movement 1, Sensation 1, Movement 2, and Sensation 2 indicate the same two-layer network over the temporal progression of the algorithm. Figure adapted with permission from Lewis et al. [15].

2.4 Classification Method

In order to enable classification of unseen objects, we extended the classification algorithm in Lewis et al. [15]. The original algorithm required the location representation of the object to be a subset of the target representation, where the target was a single learned example. In contrast, we treat inference as the location representation driving the activity of a class node via locally learned weights. Classification takes place when the activity of a class node exceeds a user-set threshold relative to the activity of other class nodes. In short, the classification step serves as a linear classifier with the active location representations as the

input features, and with weights learned via a Hebbian-like, supervised signal, rather than gradient descent. The location representation itself is determined as described in Section 2.3.

Learning During learning of the objects, whenever a particular location representation occurs (i.e. pattern of active grid cells), their identity is recorded. At the end of learning an object's local feature-location associations (described in detail in Supplementary Materials - Methods) all of the grid cells that were active during the learning of that object strengthen an associative weight with the neuron representing the class identity. This simple Hebbian-like associative learning is the only supervised signal in the system. Note that these weight updates are additive. More concretely, let the binary vector α^{loc} indicate the trace (i.e. history) of all grid cells that were active at some point during the learning of a particular object. α^{loc} is a vector of length $n^2 * m$, where n is the width of a module, and m is the number of grid cell modules. W is the associative weights learned between the grid cells and the class nodes, of which there are a total of K, one for each class, such that W is a $K \times n^2 * m$ matrix. All weight values are initialized at zero. During learning, if the object being learned is of class k, and letting j index the values of α^{loc}, then the weight updates that take place are as follows:

$$W_{kj} := \begin{cases} W_{kj} + 1, \alpha^{\text{loc}}_j = 1 \\ W_{kj}, \text{ otherwise} \end{cases} \tag{1}$$

Inference The system classifies an object once the activity for a given class is significantly higher than for the other candidate classes. Specifically, all active grid cells after a sensation, $\text{vec}(A^{\text{loc}}_{t,\text{sense}})$, feed on to the class nodes via the learned weights W. During inference, the activity vector of class nodes at time t is

$$v_t = W\text{vec}(A^{\text{loc}}_{t,\text{sense}}) \tag{2}$$

The activity of the maximally activated class node is processed via divisive normalization (a biologically plausible operation [9]). Classification takes place if this value exceeds a user set threshold γ, i.e.

$$\frac{||v_t||_\infty}{\sum_k v_{k,t}} \geq \gamma \tag{3}$$

Classification is successful if the index of the maximally active class neuron matches the true class. Finally, to prevent occasional false confidence early in inference, the system is constrained to only perform classification when $t \geq 5$, out of a total of 25 sensations.

3 Experiments

We now demonstrate the advantages of GridCellNet for operating in environments like those experienced by intelligent, biological agents. Our experiments consist of:

- The performance of a variety of classifiers, including GridCellNet, on a classification task given sequential sensory inputs. These sequences can follow arbitrary or fixed orders across training and evaluation.

- Evidence of GridCellNet's ability to rapidly develop predictive representations of visual space.

- Empirical evidence that GridCellNet's performance is dependent on integrating self-movement information with sensory data (Supplementary Material - Experiments).
- A demonstration of GridCellNet's ability to frequently perform inference with only a subset of the input sequence (Supplementary Material - Experiments).
- An evaluation of GridCellNet's robustness in a continual learning setting, in comparison to a long short-term memory (LSTM) classifier [10] (Supplementary Material - Experiments).

3.1 Translation Invariance and Inference Given Arbitrary Sequences

Vision in embodied agents, such as primates, faces the reality that a stimulus often cannot be sampled in its entirety through a single fixation, and so some mechanism must exist for integrating multiple foveal inputs. The first time an object is seen, a particular sequence of eye movements will be followed. The next time this object is observed, however, there is little-to-no guarantee that the eyes will sample it by beginning at the exact same location they did the first time it was seen, nor that they will subsequently follow the same sequence after this. As the stimulus can move in the real world, and the starting position of where classification begins can vary (i.e. translate), translation invariance represents a considerable challenge for sequential classifiers. Two requirements for translation invariance are therefore i) the ability to integrate features from arbitrary starting positions and sequence inputs, ii) the use of an internal reference frame for classification. The following section provides empirical evidence that GridCellNet satisfies (i), while it satisfies requirement (ii) by definition. In particular, no information about the absolute position of features in the external world is provided, and instead GridCellNet relies entirely on sensory inputs and self-movement information. Additionally, the following results demonstrate that GridCellNet can generalize to novel examples of MNIST that it has not seen in the training data.

For the unlikely situation of a fixed input sequence, we evaluate our classifiers given a fixed starting position and sequence of samples across the space of 5×5 features. In particular, the sequence follows the same order for all objects during both training and evaluation. With arbitrary starting positions and sequences, the input order is randomly determined for every object, and is not fixed between training and testing. As GridCellNet performs only a single weight update per feature of an object during learning, we compare to LSTMs with both 1 epoch of training as well as 50 epochs. Note therefore that our few-shot setting considers not only exposure to a limited number of training examples, but also limited opportunity for weight updates with each training example.

We begin by evaluating the classifiers on the standard task of learning and generalization given a fixed input sequence. As expected, all of the classifiers do reasonably well (Figure 2a), with the exception of the LSTM constrained to only one weight update per set of input features. GridCellNet's learning takes place using rapid Hebbian style weight updates, and so unlike the LSTM, it can form robust representations in-spite of only having observed each training object once.

We next assess performance in the setting of arbitrary starting positions and sequence inputs. As predicted, only GridCellNet maintains its performance (Figure 2b). Despite also being given self-movement information, the LSTM does not attain the same performance of GridCellNet in the few-shot setting.

GridCellNet achieves its robustness by use of the path integration properties of grid cells, enabling the network to represent the spatial location of features in a manner that can han-

dle arbitrary and previously unseen movements through space. After perceiving a feature at a given location, path integration uses self-movement to meaningfully update the active location representation. This in turn predicts the learned features at the new location. The particularities of the path that was taken is irrelevant to this process, and so the system is robust to arbitrary feature sequences.

This ability to handle arbitrary sequence paths can be viewed as a form of out-of-distribution generalization. Specifically, any particular path is astronomically unlikely to be experienced more than once. With 25 total locations, there are a total of 25! possible feature sequences, or around 15 million billion billion. Even assuming that 10 of 25 sensations is sufficient for classification, this represents $25!/15!$ possible sequences, or around 12 trillion. Thus, with any reasonable amount of training data, the classifier cannot rely on having previously observed a particular path for a particular object class.

We highlight that GridCellNet's inference operates entirely on sensory inputs and self-movement information (i.e. an internal reference frame). On the first sensation, the network initializes an internal representation based on previously learned objects, and uses subsequent movements in their reference frames to iteratively narrow down the candidate objects. As a consequence, the starting position of inference is irrelevant. If the entire stimulus was shifted in an external coordinate frame, such as the absolute location in a room, inference would be unaffected and as such, GridCellNet's classification system is an important step towards enabling online translation invariance in sequential classifiers. In practice, an end-to-end translation invariant system also requires that the actual sensory inputs (our 'foveal responses') satisfy certain levels of translation invariance and equivariance, a requirement we elaborate in further detail in the Supplementary Material (Experiments and Discussion).

Finally, we note that our same architecture, without modification, is also capable of recalling a specific learned example from the training data. This replicates the functionality of older biologically plausible models in the literature, but using an entirely internal reference frame (Supplementary Material - Experiments).

3.2 Predictive Representations

In the Supplementary Material (Experiments), we demonstrate that GridCellNet can often perform classification before the entire input sequence is experienced. This carries an additional advantage beyond efficiency. Due to the predictive nature of the network, GridCellNet can represent features in unsensed parts of the image. In particular, at any given time, GridCellNet predicts the sensory input(s) it will receive after a movement. These predictions are based on the currently active representations, which can be viewed as GridCellNet's working hypotheses about the object being sensed.

To visualise GridCellNet's representations during inference, we extract the sequence of sensed and predicted features at each progressive movement. As the sensor passes over the sequence of inputs, we accumulate the ground-truth sensations previously experienced, as well as the prediction for the next movement. Importantly however, once the system has converged to the representation of a single object, all future accumulated representations are based solely on predictions from the inputs received up until the time of convergence. The totality of these representations are then fed to a decoder network to visualise the output at different stages of inference. Additional details of this experiment are provided in the Supplementary Material (Experiments).

Figure 3 shows examples of GridCellNet restoring from memory examples that closely match the input. In the early stages of inference, the internal representation consists of a

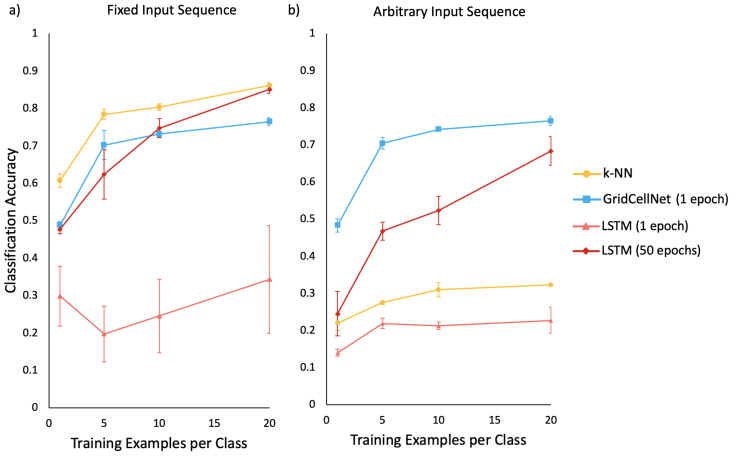

Figure 2: *Performance of Classifiers Given Fixed or Arbitrary Sequences of Input.* Classification accuracy on 1000 examples of the MNIST test set as a function of the number of training examples per-class. a) Accuracy when an identical sequence of passes over the input space is used for both training and inference. b) Performance when the starting point and subsequent sequence over the input space can be arbitrary (i.e. different) between training and inference. Error bars show the 95% confidence interval of the mean across three random seeds.

small number of previously perceived features, and the next predicted feature. While GridCellNet's prediction is only queried for a single location (centred at the yellow highlighted patch), the decoder naturally attempts to reconstruct the entire input. Furthermore, any given feature vector is an abstract representation of a 16×16 pixel region. As such, un-queried regions are not completely empty. At the point at which single-object convergence occurs, enough features have been sensed that the decoded image often appears recognizable to a human. After then evaluating predictions at every unseen location, we observe that GridCellNet can recall an example from memory that is similar to the input.

Note that although there will be overlap in the receptive fields of the different sensed inputs (each feature vector covers 25% of the input space including padding), reconstructing a detailed representation of the object requires GridCellNet to predict the abstract representation at a particular foveation, given only the abstract representations at other foveations, i.e. without direct access to the representation at a pixel-level. Despite this, one might wonder if there is sufficient data in the initial ground-truth sensations before convergence to accurately reconstruct the input. In particular, it might be possible to pad out the unsensed, predicted regions with random vectors of the same sparsity as the ground-truth features, and yet still achieve accurate reconstruction. To rule out this possibility, we provide a random predictions control, where the ground-truth features up until convergence are still provided, but predicted representations are replaced with such random vectors. These are included for the two demonstrative digits in Figure 3a, as well as for an array of additional digits in Figure 3b. As can be seen, these control reconstructions are almost entirely meaningless, and so GridCellNet's predictive functions are crucial to the plausible reconstructions observed. While some examples of GridCellNet's representations are far from perfect (in particular the '2' and '4'), the results provide evidence that these feature-level predictions could support additional downstream tasks other than classification.

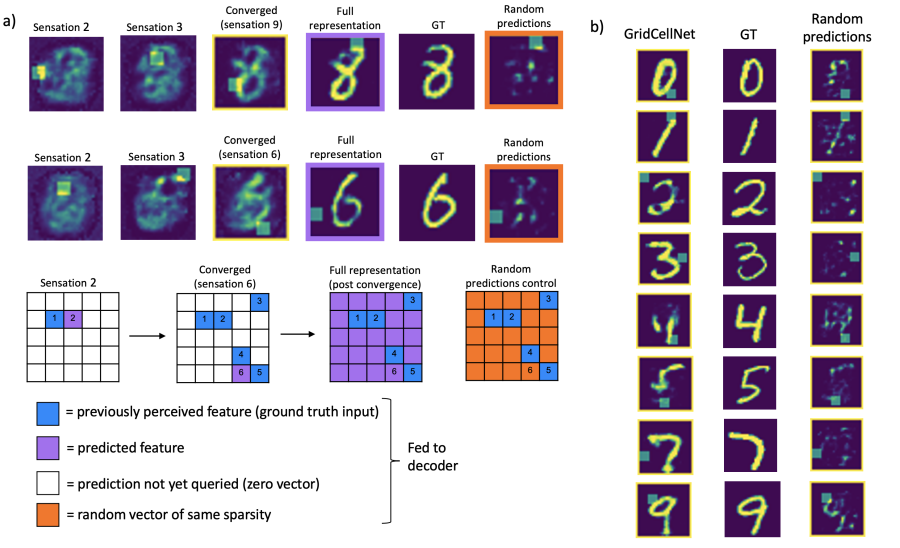

Figure 3: *Predicting Unsensed Parts of the Object.* a) We use the sensed and predicted features of GridCellNet and a decoder network to visualise the system's representations. We show two example prediction sequences (top). GridCellNet predicts its next sensation (centred at the yellow highlighted patch) based on prior input. After converging to a single representation, we visualize the system's predictions of the entire image (purple border). A diagram of the correspondence between past sensations and upcoming predictions is given at the bottom, matching the representations seen for the '6' in the middle row. At each sequence step, the entire 5×5 grid of feature representations (each a vector of length 128) is fed to a separately trained decoder. We also show a random predictions control (orange border), where predictions are replaced by random binary vectors of the same sparsity. Note "Sensation 1" is not shown, as at this time, there is no prediction for the network to make. b) Several randomly-selected examples from the other classes. GT = ground truth.

4 Conclusion

We have presented a novel approach to the challenge of visual object recognition given an arbitrary sequence of feature inputs sampled across space, a form of out-of-distribution generalization. Robustness to a novel sequence trajectory is achieved through the use of grid cells to model the location of features in the reference frame of an object. This robust classification is often rapid, occurring after only a fraction of the total input space is sampled. Finally, GridCellNet takes advantage of rapid Hebbian-style weight updates to enable few-shot and continual learning robustness, and predictive components to enable the completion of partially sensed images. Further discussion of our results and their context is provided in the Supplementary Material (Discussion).

We believe that this work supports the notion that the brain may use grid cell computations when performing visual object recognition, and that this might underlie some of the visual tasks in which humans still outperform engineered systems. Future versions of the proposed architecture that enable a complete integration of back-propagation and rapid Hebbian learning will likely be important to realising its full potential.

References

[1] Subutai Ahmad and Luiz Scheinkman. How Can We Be So Dense? The Robustness of Highly Sparse Representations. *ICML 2019 Workshop on Uncertainty and Robustness in Deep Learning*, 2019.

[2] Srdjan D. Antic, Wen Liang Zhou, Anna R. Moore, Shaina M. Short, and Katerina D. Ikonomu. The decade of the dendritic NMDA spike, 2010. ISSN 03604012.

[3] Andrej Bicanski and Neil Burgess. A Computational Model of Visual Recognition Memory via Grid Cells. *Current Biology*, 2019. ISSN 09609822. doi: 10.1016/j.cub.2019.01.077.

[4] Bjorn Browatzki, Vadim Tikhanoff, Giorgio Metta, Heinrich H. Bulthoff, and Christian Wallraven. Active in-hand object recognition on a humanoid robot. *IEEE Transactions on Robotics*, 30(5), 2014. ISSN 15523098. doi: 10.1109/TRO.2014.2328779.

[5] Ila R. Fiete, Yoram Burak, and Ted Brookings. What grid cells convey about rat location. *Journal of Neuroscience*, 28(27), 2008. ISSN 02706474. doi: 10.1523/JNEUROSCI.5684-07.2008.

[6] Torkel Hafting, Marianne Fyhn, Sturla Molden, May Britt Moser, and Edvard I. Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052), 2005. ISSN 00280836. doi: 10.1038/nature03721.

[7] Jeff Hawkins, Subutai Ahmad, and Yuwei Cui. A Theory of How Columns in the Neocortex Enable Learning the Structure of the World. *Frontiers in Neural Circuits*, 11(October):1–18, 2017. ISSN 1662-5110. doi: 10.3389/fncir.2017.00081. URL http://journal.frontiersin.org/article/10.3389/fncir.2017.00081/full.

[8] Jeff Hawkins, Marcus Lewis, Mirko Klukas, Scott Purdy, and Subutai Ahmad. A framework for intelligence and cortical function based on grid cells in the neocortex. *Frontiers in Neural Circuits*, 2019. ISSN 16625110. doi: 10.3389/fncir.2018.00121.

[9] David J. Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9(2), 1992. ISSN 14698714. doi: 10.1017/S0952523800009640.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8), 1997. ISSN 08997667. doi: 10.1162/neco.1997.9.8.1735.

[11] Joshua B. Julian, Alexandra T. Keinath, Giulia Frazzetta, and Russell A. Epstein. Human entorhinal cortex represents visual space using a boundary-anchored grid. *Nature Neuroscience*, 21(2), 2018. ISSN 15461726. doi: 10.1038/s41593-017-0049-1.

[12] Nathaniel J. Killian, Michael J. Jutras, and Elizabeth A. Buffalo. A map of visual space in the primate entorhinal cortex. *Nature*, 491(7426), 2012. ISSN 00280836. doi: 10.1038/nature11587.

[13] Hugo Larochelle and Geoffrey Hinton. Learning to combine foveal glimpses with a third-order Boltzmann machine. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010*, 2010.

[14] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2323, 1998. ISSN 00189219. doi: 10.1109/5.726791.

[15] Marcus Lewis, Scott Purdy, Subutai Ahmad, and Jeff Hawkins. Locations in the neocortex: A theory of sensorimotor object recognition using cortical grid cells. *Frontiers in Neural Circuits*, 2019. ISSN 16625110. doi: 10.3389/fncir.2019.00022.

[16] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, volume 3, 2014.

[17] Edvard I. Moser, Emilio Kropff, and May Britt Moser. Place cells, grid cells, and the brain's spatial representation system, 2008. ISSN 0147006X.

[18] Norman Mu and Justin Gilmer. MNIST-C: A Robustness Benchmark for Computer Vision. *ICML 2019 Workshop on Uncertainty and Ro- bustness in Deep Learning*, 2019.

[19] Matthias Nau, Tobias Navarro Schröder, Jacob L.S. Bellmund, and Christian F. Doeller. Hexadirectional coding of visual space in human entorhinal cortex. *Nature Neuroscience*, 21(2), 2018. ISSN 15461726. doi: 10.1038/s41593-017-0050-8.

[20] J. O'Keefe and D. H. Conway. Hippocampal place units in the freely moving rat: Why they fire where they fire. *Experimental Brain Research*, 31(4), 1978. ISSN 00144819. doi: 10.1007/BF00239813.

[21] Zachary Pezzementi, Caitlin Reyda, and Gregory D. Hager. Object mapping, recognition, and localization from tactile geometry. In *Proceedings - IEEE International Conference on Robotics and Automation*, 2011. doi: 10.1109/ICRA.2011.5980363.

[22] Marc'Aurelio Ranzato. On Learning Where To Look. *arXiv preprint arXiv:1405.5488*, 2014.

[23] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on MNIST. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[24] Marco Seeland and Patrick Mäder. Multi-view classification with convolutional neural networks, 2021. ISSN 19326203.

[25] Florian Tramèr, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, and Jorn Henrik Jacobsen. Fundamental Tradeoffs between Invariance and Sensitivity to Adversarial Perturbations. In *37th International Conference on Machine Learning, ICML 2020*, volume PartF168147-13, 2020.

[26] Alex Wong and Alan Yuille. One shot learning via compositions of meaningful patches. *Proceedings of the IEEE International Conference on Computer Vision (2015)*, 2015.

[27] Alfred L Yarbus. Eye Movements During Perception of Complex Objects, 1967.