Learning to count visual objects by combining "what" and "where" in recurrent memory

Anonymous Author(s)
Affiliation
Address
email

Abstract

Counting the number of objects in a visual scene is easy for humans but challenging for modern deep neural networks. Here we explore what makes this problem hard and study the neural computations that allow transfer of counting ability to new objects and contexts. Previous work has implicated posterior parietal cortex (PPC) in numerosity perception and in visual scene understanding more broadly. It has been proposed that action-related saccadic signals computed in PPC provide object-invariant information about the number and arrangement of scene elements, and may contribute to relational reasoning in visual displays. Here, we built a glimpsing recurrent neural network that combines gaze contents ("what") and gaze location ("where") to count the number of items in a visual array. The network successfully learns to count and generalizes to several out-of-distribution test sets, including images with novel items. Through ablations and comparison to control models, we establish the contribution of brain-inspired computational principles to this generalization ability. This work provides a proof-of-principle demonstration that a neural network that combines "what" and "where" can learn a generalizable concept of numerosity and points to a promising approach for other visual reasoning tasks.

1 Introduction

Despite significant recent advances in training deep neural networks to map between text and images, current machine learning models show significant failures in aspects of scene perception that rely on visual reasoning, such as counting how many objects are present in a natural scene [Zhang and Wu, 2020]. Even when they are equipped with inductive biases for image segmentation or salience detection, current deep supervised networks often fail to correctly enumerate objects [Lempitsky and Zisserman, 2010, Zhang et al., 2017]. Relatedly, state-of-the-art generative models such as DALLE-2 [OpenAI, 2022] struggle to craft images in response to prompts that mandate a specific numbers of objects (Fig. 1). By contrast, most children learn to count objects in the first two years of life, and it remains one of the most robust and ubiquitous human abilities [Schleifer and Landerl, 2011].

In the current paper, we make two contributions. Firstly, we use a stylised dataset to systematically define the conditions under which deep convolutional neural networks succeed and fail at generalizing their ability to count, addressing an issue that remains controversial [Wu et al., 2019]. Secondly, we describe a new model whose architecture is inspired by the parallel pathways of the primate visual system, in which visual information flows both ventrally (the "what" stream, to temporal lobe areas that represent objects in a position-invariant fashion) and dorsally (the "where" or "how" stream, to parietal lobe areas that code for the spatial position, salience and motor affordances induced by objects in a scene) [Goodale and Milner, 1992, de Haan et al., 2018, Bisley and Goldberg, 2010]. Like the primate brain, our model apprehends the scene through a series of discrete glimpses, and counting is achieved by combining both the gaze contents ("what") and gaze location ("where") in recurrent

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

memory. We use an approach first proposed by Larochelle and Hinton [2010] in which saccadic output (the movement "afforded" by the objects in the scene) is provided as an additional input to recurrent memory. The network is able to successfully count the number of items in an array and to generalise this ability to a hold-out set containing wholly novel objects. Using a symbolic model, we taxonomize counting problems according to their need for integration of "what" and "where", and show that the advantage of our dual-stream model is greatest when more integration is needed.

Our work builds on the hypothesis that dorsal and ventral streams evolved to factorize representations of the content and structure of sensory experience: the ventral stream encodes object identity while the dorsal stream encodes the abstract structure of space, events, and tasks [Summerfield et al., 2020, O'Reilly et al., 2021, Bottini and Doeller, 2020]. We propose that this factorization should enable systematic generalization (e.g., the ability to process new contents via learned structures). Convergent evidence from machine learning similarly points to the value of partitioned representations of content and structure for relational tasks [Kerg et al., 2022, van Steenkiste et al., 2019]. Our model thus integrates both ventral and dorsal stream information and we probed its ability to generalize a learned structure (numerosity) to new contents (objects and contexts).

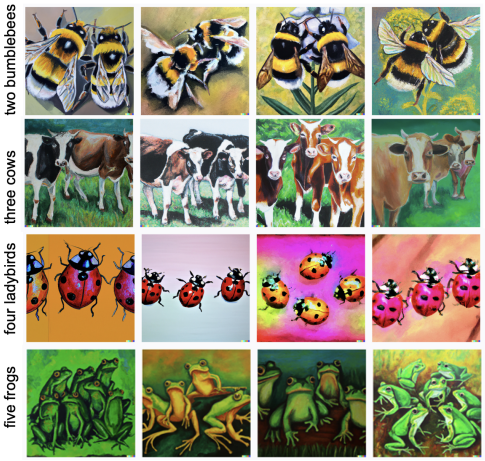

Figure 1: Images generated by DALL-E 2 in response to the prompt "A painting of . . . " followed by the text to the left if each row.

2 Related Work and Background

Visual counting in ANNs Counting and the neural code for number have been studied in neural networks using three broad approaches. Firstly, some authors have trained standard feedforward convolutional neural networks (CNNs) on annotated datasets to directly report the number of objects in a whole image, equivalent to the human ability to "subitize" or instantly grasp the number of items in an array. This works reasonably well for very small numbers but typically falters as the number of items grows [Segui et al., 2015, Zhang et al., 2017, Chattopadhyay et al., 2017, Fang et al., 2018]. Secondly, others have built models that count via explicit object detection and segmentation or density estimation [Lempitsky and Zisserman, 2010, Trott et al., 2018, Zhang et al., 2018]. These reports often use attention-based mechanisms to sequentially segment objects using a bounding box. One challenge is to disambiguate a lone object from two in close proximity [Zhang et al., 2018]. Finally, drawing on techniques and assumptions from neuroscience, some authors have sought to identify number-selective units in deep networks that were not trained on number-related objectives [Stoianov and Zorzi, 2012, Nasr et al., 2019]. However, these results have often been hard to interpret, especially given that untrained networks seemingly exhibit number neurons [Zhang and Wu, 2020, Kim et al., 2021], casting doubt on the relevance of number selectivity to visual numerosity behaviour. This, in this body of work, there is limited evidence for a general concept of numerosity.

ANN models of the dorsal stream In the wake of the well-documented correspondence between patterns of neural activity in CNNs and the primate ventral stream [Lindsay and Serre, 2021], several researchers have begun looking for a similar correspondence to the dorsal stream. Bakhtiari et al. [2021] and Mineault et al. [2021] both investigate the objective functions that yield dorsal stream-like neural activity. This line of research is related to ours by studying the purpose of the parallel pathways of the primate visual system. However, our goal is primarily to explain a cognitive ability rather than to account for neural activity. Adeli et al. [2022] tackle visual reasoning with a dual-stream, recurrent architecture which takes sequential glimpses based on an object attention mechanism. This model does not process the gaze locations, as in our work. Their model generalizes better to out-of-distribution (OOD) test examples than ResNet18, but still performs much worse on these examples than on in-distribution test images. Fabi et al. [2022] explore another dual-stream recurrent architecture for the compositional problem of learning to write characters.

Glimpsing computer vision models Larochelle and Hinton [2010] describe an architecture that takes a sequence of foveated glimpses based on a learned saccadic policy. This model learns to combine 'what' (gaze contents) and 'where' (gaze location) to solve several object recognition tasks. The Recurrent Attention model similarly processes a sequence of glimpses to obtain an efficient solution to several image classification tasks [Mnih et al., 2014]. Inspired by human-like iterative and attentive counting processes, [Ren and Zemel, 2017] propose a recurrent architecture with visual attention for object instance segmentation. They jointly solve the counting and instance segmentation problems, segmenting one instance at a time.

Visual numerosity in the primate brain In the primate brain, mounting evidence implicates the posterior parietal cortex (PPC) in visual numerosity and in number sense more generally [Nieder and Dehaene, 2009]. Monkey electrophysiology has found accumulator neurons as well as neurons tuned to specific numerosities in intraparietal regions [Nieder and Miller, 2004, Roitman et al., 2007]. Human neuroimaging tells a similar story, reporting regions tuned to visual numerosity [Piazza et al., 2004] that are arranged topographically in PPC [Harvey et al., 2013, 2015, Harvey and Dumoulin, 2017, Cai et al., 2021]. These neural correlates are also consistent with lesion studies in which patients with damage to parietal cortex show deficits in numerical cognition, including visual tasks like dot counting [Ashkenazi et al., 2008, Takayama et al., 1994].

3 A Neuro-Inspired Recurrent Glimpse Network

3.1 The Counting Task

We synthesize grayscale images (27 x 21 pixels) each containing 2-6 items (alphanumeric characters). Items are of constant size and occupy a maximum area of 5 x 3 pixels, and are placed on a uniform background of contrasting luminance. We divide each image into a 3 x 3 grid in which each grid square is a possible item location. We pose the counting task as a supervised classification problem, minimizing a cross-entropy loss between the actual and predicted number of items. During training, foreground and background luminances are sampled randomly from the set $\{0, 0.5, 1\}$ and item shapes are randomly chosen to be one of the digits 0–4. Model performance is evaluated on a validation set consisting of new images from within the training distribution and three OOD test sets consisting of new shapes (the digits 5–9), new luminances (intermediate grays $\{0.1, 0.3, 0.7, 0.9\}$) and both new shapes and new luminances. In our main experiments, all items within an image are the same (homogenous sets) but see Appendix for comparable results with varied heterogenous sets.

3.2 Modeling saccadic trajectories

We call our model the Recurrent Glimpse Net (RGN). Rather than process the entire image in parallel, the RGN receives a sequence of partial glimpses (6 x 6 pixels each) from the image. We used a fixed policy for sampling glimpse locations that obeys the following constraints: (i) there are always six glimpses (the maximum possible number of items); (ii) each glimpse location is drawn from a truncated Gaussian distribution peaking centrally within a grid square containing an item (s.d. ≈ 2 pixels, truncation at ≈ 4 pixels); and (iii) all items are glimpsed at least once. Whilst this policy is stylised, it approximates that which is obtained by sampling from a visual "salience map", and it replicates many aspects of saccadic behaviour. For example, the Gaussian is sufficiently broad that glimpses frequently span multiple items (or the edges of multiple items), as in Fig. 2b.

3.3 Dual Stream Processing

Just as the primate visual system processes "what" and "where", the saccadic glimpsing produces two streams of observations. One contains the sequence of extracted pixels (the 6 x 6 contents of each glimpse) and the other contains the sequence of gaze locations (the [x,y] Cartesian coordinates of the centre of each glimpse). In our task, as in natural vision, neither stream permits accurate counting on its own. Since all items within an image are identical, numerosity information in the gaze contents stream is highly ambiguous. From one glimpse to the next, it is not obvious whether an item (or item fragment) is one that has been previously glimpsed (in which case the count should not be incremented) or is new (in which case it should). Similarly, numerosity is ambiguous from the the gaze location stream alone. For two nearby gaze locations, it is unclear whether the eye is directed towards the same item twice, or two different items.. However, our work tests the contention that

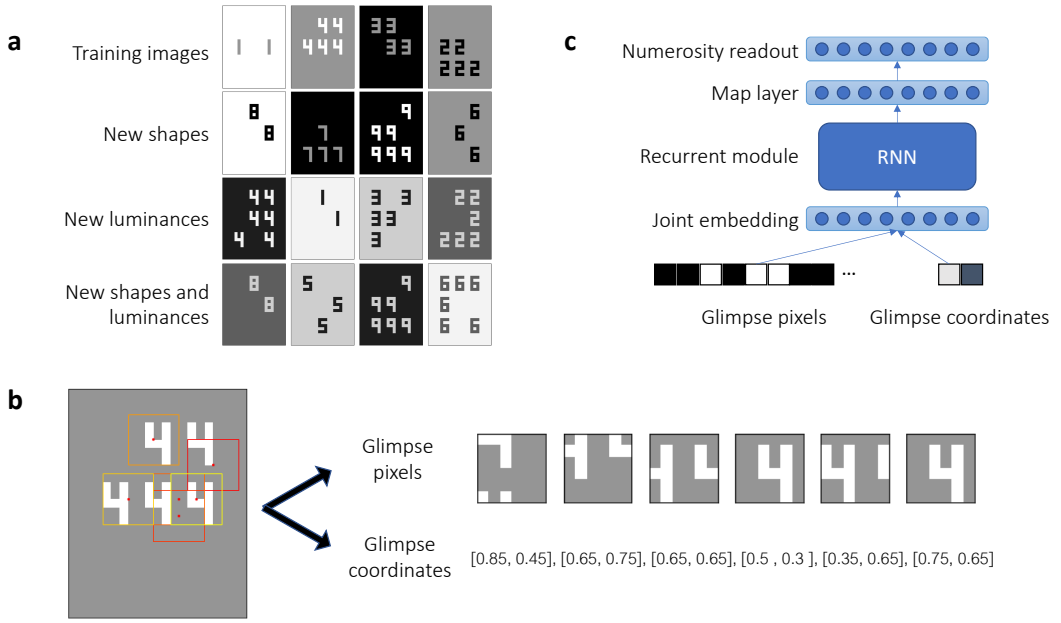

Figure 2: Task and model description. (a) All images contain 2-6 items. Models are trained on a subset of item shapes and luminances and tested on held out image parameters. (b) Each image is glimpsed, producing two sequences of observations: the glimpse pixels (gaze contents) and the glimpse coordinates (gaze location). (c) The Recurrent Glimpse Net processes the two glimpse streams to produce a spatial map of the items in the image from which a numerosity label is computed.

these two sources of ambiguity can be resolved by combining information from both streams, and that numerosity judgements will require an integration of "what" and "where".

3.4 Network Architecture

The input layer of the network consists of 36 units for the flattened glimpse contents plus two additional units for the gaze locations. These feed into a joint embedding layer preceding a recurrent module. The penultimate layer of processing before the numerosity readout is trained to represent a spatial map of the items in the image via an auxiliary loss. This layer consists of nine units corresponding to the nine possible item locations. The map loss term is a binary cross entropy loss whose target is whether the image contains an item in each of the nine 'slots'. The full architecture is depicted in Fig. 2 and detailed network and training parameters can be found in the Appendix. All models and experiments are coded in Python 3.7.5 using PyTorch 1.11.0.

4 Experiments

All models were trained with stochastic gradient descent on 100,000 training images for 500 epochs. Each of the four test sets consisted of 5,000 images. All results are averaged over 10 random initializations. Detailed training parameters can be found in the Appendix.

4.1 Recurrent Glimpse Net generalizes to new shapes and new luminances

We compare our RGN to a parameter-matched convolutional neural network that receives the entire image as input rather than a sequence of glimpses. The learning curves in Fig. 3 show the performance on the training set and the four test sets. We find that our RGN (top row) is able to generalize to new shapes and new luminances, achieving $> 96\%$ accuracy on all test sets. The CNN (bottom row) is able to learn the task but suffers considerably on all the OOD test sets.

The remaining analyses are designed to probe how the RGN is able to generalize. Here we are verifying whether the neuroscience- and cognitively-inspired aspects of the model are bearing their expected fruits and assigning credit to the various computational ingredients.

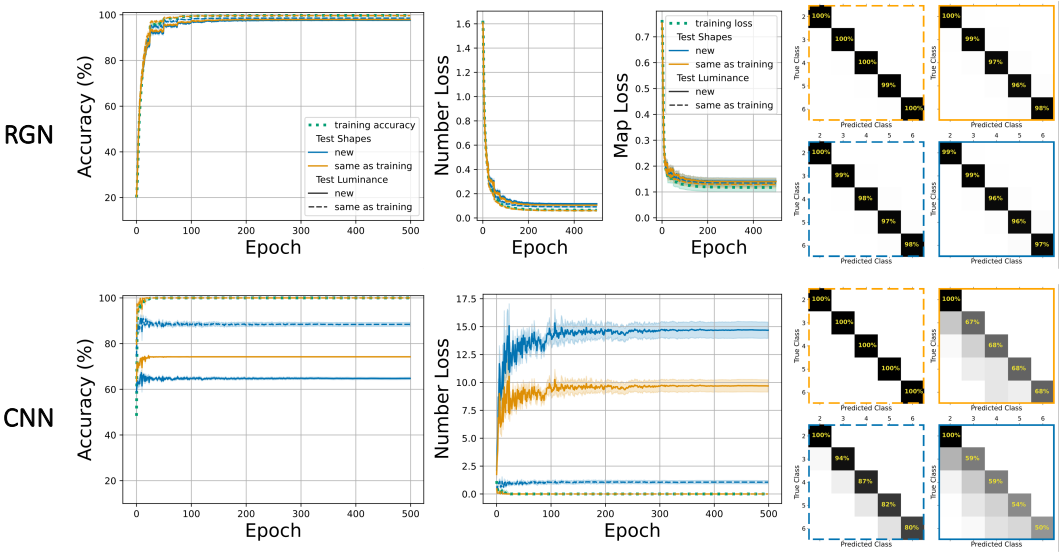

Figure 3: (top) Our Recurrent Glimpse Net generalizes to new shapes and new luminances. (bottom) A parameter matched CNN shows significant deficits on all of the OOD test sets. Accuracy and loss curves show mean and SEM over 10 random initializations.

4.2 Recurrent Glimpse Net performance depends on the integration of both input streams

The architecture of the RGN is inspired by the premise that numerosity judgements require the integration of "what" and "where". However, the need for integration varies from image to image, and depends on the precise configuration of items and sampled gaze locations. For some images, the gaze location stream alone could be sufficient to determine the numerosity[1]. We wanted to see how the performance of the RGN (and control models) depends on the need for integration.

To this end, we developed a manual algorithm that derives numerosity from a symbolic representation of the glimpses. This symbolic model first tries to determine the numerosity from the gaze locations alone, and if successful, assigns the glimpsed image an integration score of 1, indicating that no integration is required. Otherwise, the counter iteratively queries a symbolic form of the glimpse contents to resolve the ambiguity about where items are located. The more times the counter needs to query the glimpse contents, the higher the integration score. The symbolic counter successfully determines the numerosity of 97.24% of the test images. The remaining 2.76% correspond to a small number of edge cases the symbolic model cannot handle and are omitted from the presented analysis. Full details of the symbolic model can be found in the Appendix.

In Fig. 4, we plot performance of three versions of the RGN on the test set consisting of both new shapes and new luminances, split by integration score. Version 1 (left) receives only the gaze locations as input, version 2 (middle) receives only the glimpse contents as input, and version 3 (right) receives both input streams. Neither of the single stream models manages to master the task. As-expected, performance of the locations-only model scales with integration score, achieving 100% accuracy on only those images with an integration score of 1. Only the model with both input streams performs well on all integration scores by the end of training. It masters the 'easier' images faster, requiring more training for the higher integration score images. This confirms that RGN's generalization performance relies on an integration of both input streams, and supports our theory that enactive inputs (gaze locations) can help solve visual reasoning tasks by providing information about the relational geometry among items.

[1]Although not present in the datasets analyzed here, there could also be cases where the pixel stream alone is sufficient to determine the numerosity, for example, if all items in an image were distinct and the number of items was equal to the maximum numerosity in the dataset.

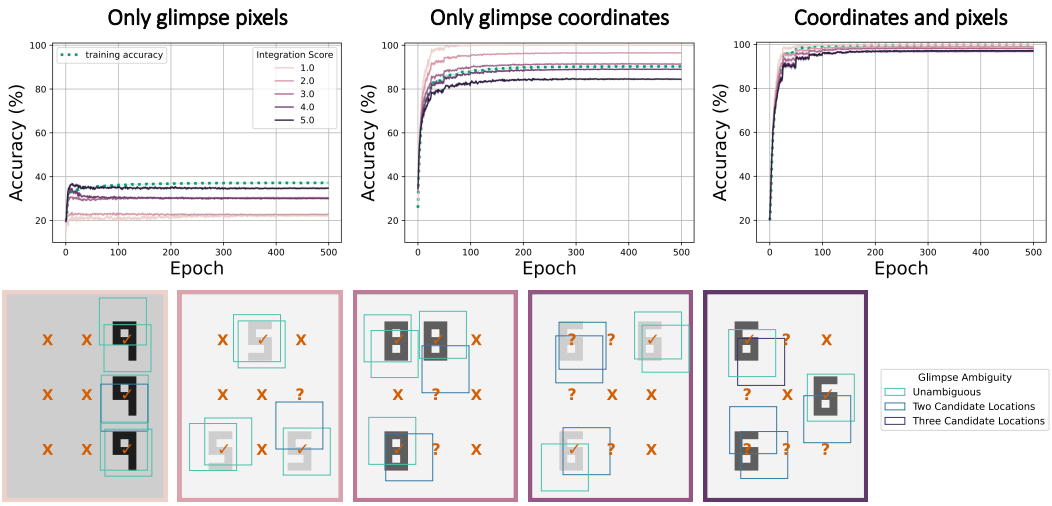

Figure 4: (top) Generalization performance on the new shapes and new luminances test set for the RGN which receives only the gaze contents (top left), only the gaze locations (top middle), or both input streams (top right) as a function of integration score. Including both input stream closes the gap between integration scores observed with locations alone. Training performance is included for reference (cyan dashed line). (bottom) Illustration of the symbolic model. Each panel shows an example image with glimpses (blue boxes, coloured to denote ambiguity about item-location assignments) overlaid on items. Panels are arranged from left to right in increasing order of integration score (1-5). The orange ✓, X, and ? symbols indicate spatial locations where the counter respectively determined that there *is*, *is not*, or *could be* an item from the gaze locations alone. For an integration score of 1 (far left), there are no interrogatives because there is no ambiguity about where the items lie.

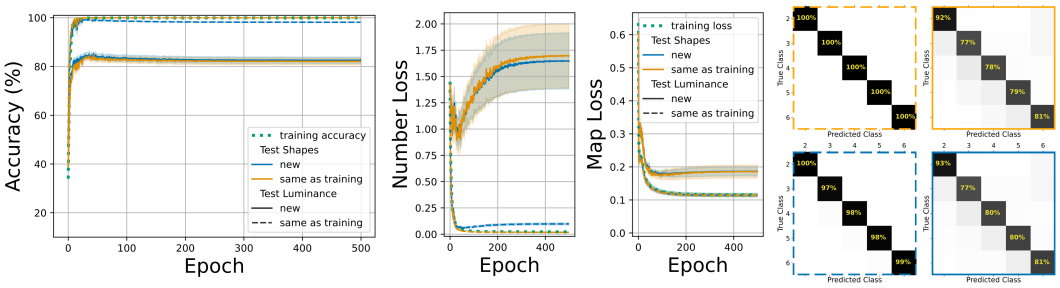

Figure 5: Performance of a non-glimpsing recurrent control model of identical architecture as RGN. Without the glimpse coordinate input, this model fails to generalize to new luminances.

4.3 Generalization to new luminances depends on the glimpse coordinate input

Could it be that recurrence alone is sufficent for visual counting in our task? To address this, we use a matched recurrent architecture that, instead of a sequence of glimpses, receives the whole image as input repeatedly. Thus, it lacks the ambiguity in the pixel stream introduced by the partial glimpses, but is also missing the glimpse coordinate input present in RGN. This recurrent control fares better than the CNN on the new shapes test set, but still fails to generalize to new luminances (Fig. 5). This suggests that the glimpse coordinate input is crucial for learning a representation of numerosity that is invariant to luminance.

4.4 A representation of space aids robust generalization in RGN

In a subsequent analysis we sought to understand the contribution of the spatial map in the penultimate layer. The symbolic model relies heavily on building an approximate spatial map of item locations, i.e. to "fill in" whether an item is present at each candidate location. Notably, biological brains compute spatial maps in the dorsal stream, although their exact computational role is not clear. We thus asked whether this map representation is necessary and/or sufficient for generalizable numerosity

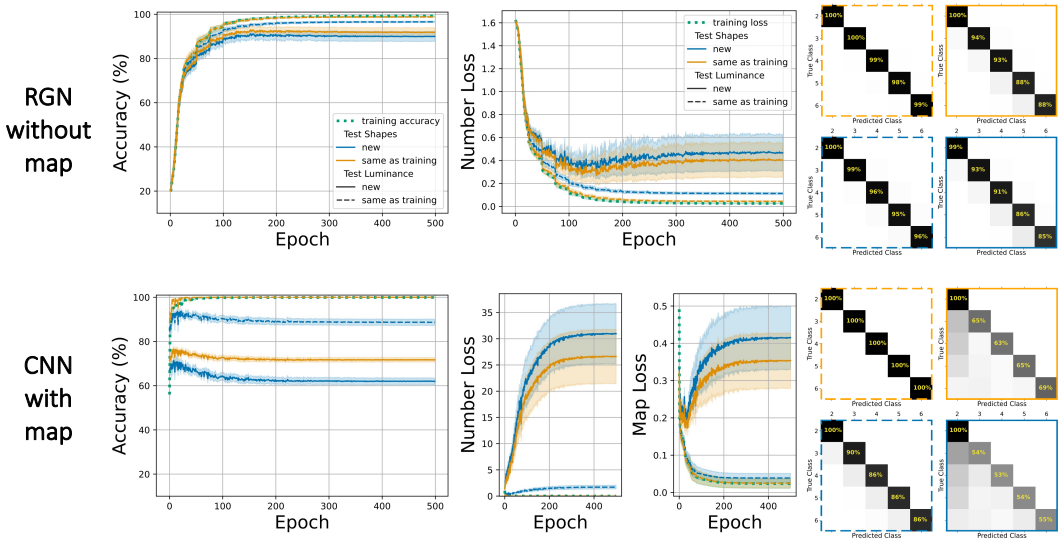

Figure 6: (top) Removing the map objective slows learning and impairs generalization in RGN. (bottom) Adding the map loss term to the CNN did not improve generalization.

judgements by training a version of RGN without the map objective and a version of the CNN baseline with the auxiliary map objective.

Firstly, we observe that adding the map objective to the CNN baseline was insufficient to render its predictions generalizable. This implies that the map loss is not sufficient for OOD counting. However, removing the map loss term from the RGN objective impairs generalization ability, especially for new luminances. This version of the model learns slower and is more sensitive to random initialization. This implies that the map loss is necessary for accurate counting under the conditions studied here. In the model, the spatial map may help the network organise relational information about the scene (what goes where), which in turn facilitates inferences, such as how many items are present. More generally, these two computational ingredients, the dual stream glimpsing architecture and the spatial map, work together in RGN to learn a generalizable concept of numerosity.

4.5 Summary

Fig. 7 summarizes generalization performance on the hardest test set (new shapes and new luminances) for all tested models as a function of the input the model received and whether the map loss term was included in the optimized objective function. The models that generalize best are those that receive the gaze locations as input, but the gaze locations on their own are not sufficient for perfect generalization. When gaze locations are combined with the gaze contents, models can generalize well, but only those models trained with the auxiliary map objective generalize well *consistently*. Without the map objective, generalization performance varies from run to run. Models trained on gaze contents alone perform only slightly better than chance. Including the auxiliary map objective was not sufficient to enable generalization in models that received the entire image as input.

5 Discussion

We designed RGN to embody computational principles hypothesized to contribute to visual scene understanding in the primate brain. We then applied RGN in an idealized and controlled task setting which allowed for careful inspection of generalization behaviour. Our goal was not to train a network to count under naturalistic conditions but to use careful, controlled experiments to study the factors that make counting hard and the computational mechanisms it may require. We compared RGN to several parameter-matched convolutional and recurrent architectures to explore the factors contributing to the RGN's generalization ability, and consequently to provide support for a theory of visual scene understanding.

The critical feature of the RGN is that it combines gaze locations ("where") with gaze contents ("what"). The gaze locations stand in for what could in biology be an overt attention-driven movement, as in a saccade, or a covert attentional signal in the absence of an explicit movement. In the setting we study, this signal alone is very useful for computing numerosity, but fully accurate performance on new, held-out displays is only possible by integrating information from both streams. We note the caveat that here, we only study a single stylised task under a fixed (if broadly plausible) saccadic policy. More work is required to establish how sensitive our findings are to the specific conditions (e.g., the presence of clutter) and saccadic policy.

Our saccadic policy was designed to yield glimpse sequences that spanned our integration score. We ensured that the vast majority of glimpse sequences received an integration score of 2 or higher, meaning that our symbolic counter required querying both input streams to determine the numerosity. This dataset design allowed us to quantify the specific deficits associated with unistream models. We showed that the performance of a model receiving only gaze locations grades with integration score, generalizing perfectly on glimpse sequences with an integration score of 1 (no integration required). This result validated our interpretation of RGN's generalization success as the result of integrating both input streams. However, it remains to be seen how biologically realistic these saccadic trajectories are. In future work, we plan to explore more biologically plausible models of salience-based attention and to collect eye-tracking while human subjects view images similar to those used here. Our theory will be further validated if we can predict human visual reasoning responses from real glimpse sequences.

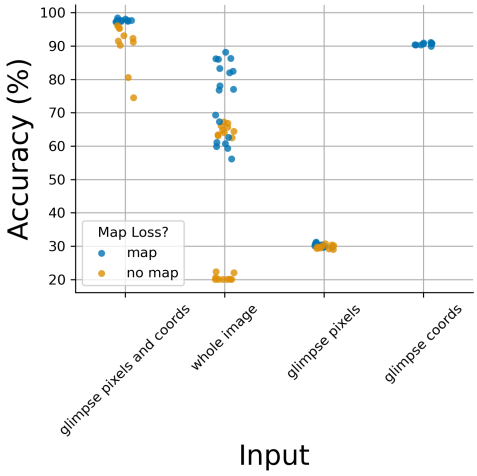

Figure 7: Accuracy on the hardest test set as a function of input and whether the auxiliary map objective was optimized. Each point is one model run.

Acquiring the abstract concept of numerosity was one of the Bongard Problems for AI posed over 50 years ago [Bongard, 1970], alongside other relational problems like same-different tasks that still challenge modern AI systems [Stabinger et al., 2021, Mitchell, 2021, Chollet, 2019]. In the present work, numerosity recognition is primarily a spatial-relation task. In future work, we plan to extend our approach to other tasks involving an integration of multiple object relations and identities, for example, counting objects that satisfy some query in cluttered images, comparing quantities of different object types, and recognizing Gestalt properties. Such tasks will invariably require a more developed ventral stream, but we expect the computational principles studied here to transfer well to other task settings.

References

Hossein Adeli, Seoyoung Ahn, and Gregory Zelinsky. A brain-inspired object-based attention network for multi-object recognition and visual reasoning. *bioRxiv*, pages 1–25, 2022.

Sarit Ashkenazi, Avishai Henik, Gal Ifergane, and Ilan Shelef. Basic numerical processing in left intraparietal sulcus (IPS) acalculia. *Cortex*, 44(4):439–448, 2008. ISSN 00109452. doi: 10.1016/j.cortex.2007.08.008.

Shahab Bakhtiari, Patrick Mineault, Tim Lillicrap, Christopher C Pack, and Blake A Richards. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. In *NeurIPS*, page 2021.06.18.448989, 2021. doi: https://doi.org/10.1101/2021.06.18.448989.

James W. Bisley and Michael E. Goldberg. Attention, intention, and priority in the parietal lobe. *Annual Review of Neuroscience*, 33:1–21, 2010. ISSN 0147006X. doi: 10.1146/annurev-neuro-060909-152823.

Mikhail Moiseevitch Bongard. *Pattern Recognition*. Hayden Book Co., Spartan Books, Rochelle Park, N.J., 1970.

Roberto Bottini and Christian F. Doeller. Knowledge Across Reference Frames: Cognitive Maps and Image Spaces. *Trends in Cognitive Sciences*, 24(8):606–619, 2020. ISSN 1879307X. doi: 10.1016/j.tics.2020.05.008. URL `https://doi.org/10.1016/j.tics.2020.05.008`.

Yuxuan Cai, Shir Hofstetter, Jelle van Dijk, Wietske Zuiderbaan, Wietske van der Zwaag, Ben M. Harvey, and Serge O. Dumoulin. Topographic numerosity maps cover subitizing and estimation ranges. *Nature Communications*, 12(1):1–10, 2021. ISSN 20411723. doi: 10.1038/s41467-021-23785-7. URL `http://dx.doi.org/10.1038/s41467-021-23785-7`.

Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R. Selvaraju, Dhruv Batra, and Devi Parikh. Counting everyday objects in everyday scenes. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 4428–4437, 2017. doi: 10.1109/CVPR.2017.471.

François Chollet. On the Measure of Intelligence. *arXiv:1911.01547v2*, 2019.

Edward H.F. de Haan, Stephen R. Jackson, and Thomas Schenk. Where are we now with 'What' and 'How'? *Cortex*, 98:1–7, 2018. ISSN 19738102. doi: 10.1016/j.cortex.2017.12.001. URL `https://doi.org/10.1016/j.cortex.2017.12.001`.

Sarah Fabi, Lena Holzwarth, and Martin V. Butz. Efficient learning through compositionality in a CNN-RNN model consisting of a bottom-up and a top-down pathway. In *Proceedings of the Annual Meeting of the Cognitive Science*, 2022.

M. Fang, Z. Zhou, S. Chen, and J. L. McClelland. Can a recurrent neural network learn to count things? *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 360–365, 2018.

Melvyn A. Goodale and A. David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992. ISSN 09758585. doi: 10.1016/0166-2236(92)90344-8.

B. M. Harvey, B. P. Klein, N. Petridou, and S. O. Dumoulin. Topographic Representation of Numerosity in the Human Parietal Cortex. *Science*, 341(September):1123–1126, 2013. doi: 10.4159/harvard.9780674333987.c22.

Ben M. Harvey and Serge O. Dumoulin. A network of topographic numerosity maps in human association cortex. *Nature Human Behaviour*, 1(2), 2017. ISSN 23973374. doi: 10.1038/s41562-016-0036.

Ben M. Harvey, Alessio Fracasso, Natalia Petridou, and Serge O. Dumoulin. Topographic representations of object size and relationships with numerosity reveal generalized quantity processing in human parietal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 112(44):13525–13530, 2015. ISSN 10916490. doi: 10.1073/pnas.1515414112.

Giancarlo Kerg, Sarthak Mittal, David Rolnick, Yoshua Bengio, Blake Richards, and Guillaume Lajoie. On Neural Architecture Inductive Biases for Relational Tasks. pages 1–23, 2022. URL `http://arxiv.org/abs/2206.05056`.

Gwangsu Kim, Jaeson Jang, Seungdae Baek, Min Song, and Se Bum Paik. Visual number sense in untrained deep neural networks. *Science Advances*, 7(1):1–9, 2021. ISSN 23752548. doi: 10.1126/sciadv.abd6127.

Hugo Larochelle and Geoffrey Hinton. Learning to combine foveal glimpses with a third-order Boltzmann machine. In *Neural Information Processing Systems*, 2010. doi: 10.1016/S0196-9781(02)00145-6.

Victor Lempitsky and Andrew Zisserman. Learning To Count Objects in Images. In *Neural Information Processing Systems*, 2010.

Grace W Lindsay and Thomas Serre. Deep Learning Networks and Visual Perception. In *Oxford Research Encyclopedia of Psychology*. 2021. ISBN 9780190236557. doi: https://doi.org/10.1093/acrefore/9780190236557.013.841.

Patrick J Mineault, Shahab Bakhtiari, Blake A Richards, and Christopher C Pack. Your head is there to move you around: Goal-driven models of the primate dorsal pathway. *bioRxiv*, (NeurIPS): 2021.07.09.451701, 2021.

Melanie Mitchell. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 2021. ISSN 0077-8923. doi: 10.1111/nyas.14619.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. *Advances in Neural Information Processing Systems*, pages 2204–2212, 2014. ISSN 10495258.

Khaled Nasr, Pooja Viswanathan, and Andreas Nieder. Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Science Advances*, 5(5):1–11, 2019. ISSN 23752548. doi: 10.1126/sciadv.aav7903.

Andreas Nieder and Stanislas Dehaene. Representation of number in the brain. *Annual Review of Neuroscience*, 32:185–208, 2009. ISSN 0147006X. doi: 10.1146/annurev.neuro.051508.135550.

Andreas Nieder and Earl K. Miller. A parieto-frontal network for visual numerical information in the monkey. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (19):7457–7462, 2004. ISSN 00278424. doi: 10.1073/pnas.0402239101.

OpenAI. DALLE-2, 2022. URL `https://openai.com/dall-e-2/`.

Randall C. O'Reilly, Charan Ranganath, and Jacob L. Russin. The Structure of Systematicity in the Brain. pages 1–8, 2021. ISSN 0963-7214. doi: 10.1177/09637214211049233.

Manuela Piazza, Véronique Izard, Philippe Pinel, Denis Le Bihan, and Stanislas Dehaene. Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, 44(3):547–555, 2004. ISSN 08966273. doi: 10.1016/j.neuron.2004.10.014.

Mengye Ren and Richard S. Zemel. End-to-end instance segmentation with recurrent attention. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:293–301, 2017. doi: 10.1109/CVPR.2017.39.

Jamie D. Roitman, Elizabeth M. Brannon, and Michael L. Platt. Monotonic coding of numerosity in macaque lateral intraparietal area. *PLoS Biology*, 5(8):1672–1682, 2007. ISSN 15449173. doi: 10.1371/journal.pbio.0050208.

Patrick Schleifer and Karin Landerl. Subitizing and counting in typical and atypical development. *Developmental Science*, 14(2):280–291, 2011. ISSN 1363755X. doi: 10.1111/j.1467-7687.2010.00976.x.

Santi Segui, Oriol Pujol, and Jordi Vitria. Learning to count with deep object features. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2015-Octob:90–96, 2015. ISSN 21607516. doi: 10.1109/CVPRW.2015.7301276.

Sebastian Stabinger, David Peer, Justus Piater, and Antonio Rodríguez-Sánchez. Evaluating the progress of deep learning for visual relational concepts. *Journal of Vision*, 21(11):11–23, 2021. ISSN 15347362. doi: 10.1167/jov.21.11.8.

Ivilin Stoianov and Marco Zorzi. Emergence of a 'visual number sense' in hierarchical generative models. *Nature Neuroscience*, 15(2):194–196, 2012. ISSN 10976256. doi: 10.1038/nn.2996.

Christopher Summerfield, Fabrice Luyckx, and Hannah Sheahan. Structure learning and the posterior parietal cortex. *Progress in Neurobiology*, 184(September 2019), 2020. ISSN 18735118. doi: 10.1016/j.pneurobio.2019.101717.

Yoshihiro Takayama, Morihiro Sugishita, Ichiro Akiguchi, and Jun Kimura. Isolated Acalculia due to Left Parietal Lesion. *Archives of Neurology*, 51(3):286–291, 1994.

Alexander Trott, Caiming Xiong, and Richard Socher. Interpretable counting for visual question answering. *International Conference on Learning Representations*, 2018.

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? *Advances in Neural Information Processing Systems*, 32, 2019. ISSN 10495258.

Xiaolin Wu, Xi Zhang, and Xiao Shu. Cognitive deficit of deep learning in numerosity. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pages 1303–1310, 2019. ISSN 2159-5399. doi: 10.1609/aaai.v33i01.33011303.

Jianming Zhang, Shugao Ma, Mehrnoosh Sameki, Stan Sclaroff, Margrit Betke, Zhe Lin, Xiaohui Shen, Brian Price, and Radomír Měch. Salient Object Subitizing. *International Journal of Computer Vision*, 124(2):169–186, 2017. ISSN 15731405. doi: 10.1007/s11263-017-1011-0.

Xi Zhang and Xiaolin Wu. On numerosity of deep neural networks. *Advances in Neural Information Processing Systems*, 2020. ISSN 10495258.

Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, pages 1–17, 2018.

A Appendix

A.1 Additional Controls

To verify that RGN's generalization ability is not limited to the case of homogenous item sets, we also trained the same model on images with heterogenous item sets. Here the shapes within an image are selected randomly. In comparing Fig. A1 and Fig. A2, we see that the gaze contents are more informative during training when the items are not all the same than when they are, but in a way that does not generalize well to the OOD test sets. The model that receives both inputs still learns to integrate both streams and generalize to new shapes, but struggles somewhat with new luminances compared to when all items are the same, presumably because it has learned to rely more on the gaze contents.

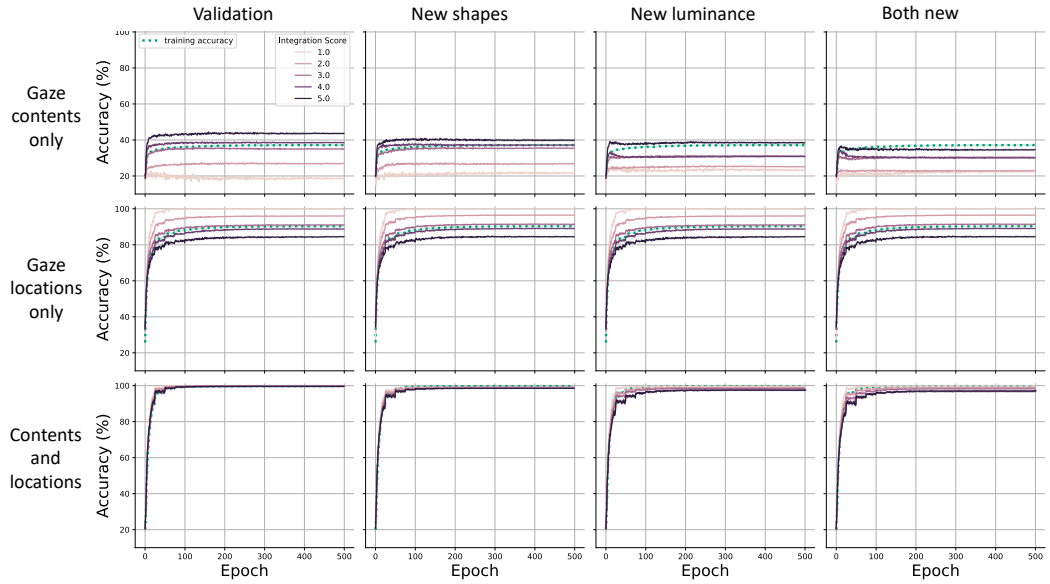

Figure A1: Accuracy of uni-stream and dual-stream RGN versions for homogenous item sets divided by integration score.

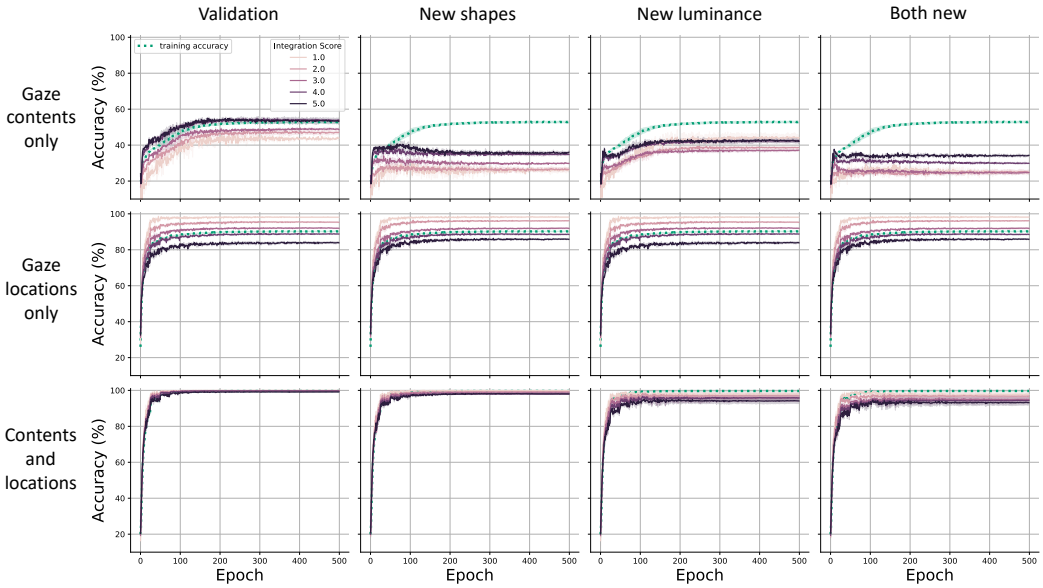

Figure A2: Accuracy of uni-stream and dual-stream RGN versions for heterogenous item sets divided by integration score.

A.2 Model and Training Specifications

All models were initialized with random weights and trained for 500 epochs with stochastic gradient descent. We used a momentum factor of 0.9 and a step learning rate schedule, decreasing by a factor of 0.7 every 25 epochs. Dropout (p=50%) was applied before the map layer in all models during training. The CNN included a 2x2 max-pooling layer before the second conv layer. A sigmoid nonlinearity was applied to the map layer units. All other units (except for the number readout units) were leaky rectified linear units with a negative slope of 0.1.

Table A1: RGN trainable model parameters

Layer	Shape	Parameters
embedding	[100, 38]	3800
i2h	[100, 200]	20000
h2o	[100, 100]	10000
map	[9, 100]	900
number readout	[7, 9]	63
Total params with biases		35079

Table A2: CNN trainable model parameters

Layer	Shape	Parameters
conv1	[33, 1, 3, 3]	287
conv2	[33, 33, 2, 2]	4356
conv3	[33, 33, 2, 2]	4356
fc1/map	[9, 2904]	26136
number readout	[7, 9]	63
Total params with biases		35323

A.3 Symbolic Model of Counting

Here we describe an idealization of the problem of counting the number of objects in an image.

A.3.1 Data synthesis

An image consists of a 1x1 square containing $2 \geq N_o \leq 6$ objects. Each object O_i for $i = 1 \ldots N_o$ is described by its Cartesian coordinates $\mathbf{c_o}_i \in \{(x_i, y_i) | x_i \in S \wedge y_i \in S\}$ where $S = \{0.2, 0.5, 0.8\}$ and its shape. $\mathbf{s_o}_i$ is a one-hot vector indicating which of N_s shapes O_i is. This defines a 3x3 grid within the image where objects may lie. These nine locations are labeled $0 \ldots 8$. So we can indicate the set of all possible object locations $A = \{0, 1, 2, 3, 4, 5, 6, 8\}$. For a given image, the set of actual object locations L will be a subset $L \subset A$ where $|L| = N_o$, i.e., the number of 'filled' locations is equal to the image's numerosity.

We then simulate taking a sequence of N_g noisy glimpses of this image. Similarly to the objects, each glimpse G_j is described by its Cartesian coordinates $\mathbf{c_g}_j$ and a shape feature vector $\mathbf{s_g}_j$ for $j = 0 \ldots N_g$. To generate the gaze locations, we first construct a random sequence of image objects, making sure that the sequence contains every object in the image at least once. For example, in an image that contains 3 objects, our sequence might be O_2, O_1, O_3, O_3. Therefore, there is some map $f(j)$ that maps from the glimpse index to the object index. The corresponding sequence of gaze locations will be the coordinates of those objects plus some noise: $\mathbf{c_g}_j = \mathbf{c_o}_{f(j)} + \epsilon_j$. The noise for x and y are sampled independently: $\epsilon_{jx} \sim \mathcal{N}(x_{f(j)}, \sigma^2)$ and $\epsilon_{jy} \sim \mathcal{N}(y_{f(j)}, \sigma^2)$. In this version, the noise is also bounded such that the total euclidean distance between $\mathbf{c_g}_j$ and $\mathbf{c_o}_{f(j)}$ does not exceed some threshold $g(\sigma^2)$, but this could be relaxed in a fully Bayesian version of the counter. The glimpse shape feature $\mathbf{s}_j^{(g)}$ is also constructed from the object shape features. Whereas the mth element of $\mathbf{s_o}_i$ indicates whether object i is an instance of shape m, the mth element of $\mathbf{s_g_j}$ is the proximity of glimpse j to objects of shape m.

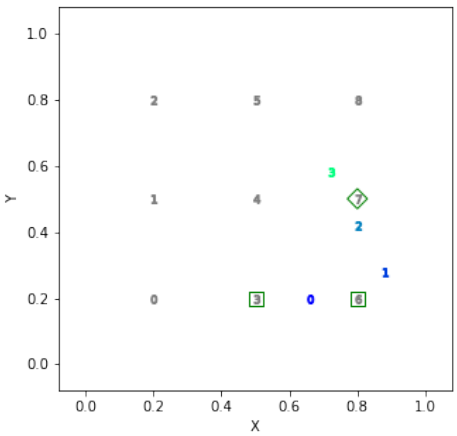

Figure A3: An idealized image depicting the nine possible object locations. In this example, the image contains three objects: a square at location 3, another square at location 6, and a diamond at location 7. Glimpse locations are indicated by the blue-green numerals.

$$\mathbf{s_g}_j^{(m)} = \begin{cases} \sum_{k \in K} 1 - \frac{dist(\mathbf{c_g}_j, \mathbf{c_o}_k)}{g(\sigma^2)} & \text{if } |K| > 0 \\ 0 & \text{otherwise} \end{cases}$$

(1)

where K is the set of object indices for which $\mathbf{s_o}_k^{(m)} = 1$ and $dist(\mathbf{c_g}_j, \mathbf{c_o}_k) \leq g(\sigma^2)$. This proximity score ranges from 0 to 1 where 0 indicates that there are no objects of shape m in the vicinity of that glimpse and 1 could mean either that the gaze locations are equal to the coordinates of an object of shape m, or that a glimpse is in between two objects of shape m. If the glimpse is close to more than two objects of the same shape, this proximity value can exceed 1.

A.3.2 Counter

Given only a sequence of glimpses $G_j = (\mathbf{c_g}_j, \mathbf{s_g}_j)$ for $j = 1 \ldots N_g$, the noise level σ^2 with which the glimpses with synthesized, and knowledge of the minimum and maximum numerosity, the counter tries to determine the number of objects in the image. It will do so by estimating L, the set of

object locations, or taking a shortcut based on prior knowledge about the minimum and maximum possible number of objects, if available. The counter will assign each image a difficulty level ranging from 0–6 which indicates the degree to which the two input streams need to be integrated in order to determine the numerosity.

Shortcuts There are two shortcuts that might be available, one based on s_g and one based on c_g. The number of unique shapes represented in s_g provides a lower bound on the number of objects in the image. Our lower bound is initialized to the minimum numerosity. If the number of unique shapes is greater than this initial lower bound, we will update the lower bound to be equal to the number of unique shapes. When the number of unique shapes is equal to the maximum numerosity present in the dataset, the numerosity can be predicted exactly without further processing. Images for which this shortcut is available are assigned a difficulty level of 0. If this shortcut is unavailable, the next step is to process c_g.

Processing gaze locations For each glimpse coordinate tuple c_g, we infer a set of candidate object locations. This will be the set of object locations within some radius (defined by the noise level) of the glimpse coordinate. For example, in Figure 2, the set of candidate locations for G_0 would be $3, 6$; G_0 indicates that there could be an object in location 3, in location 6, or in both location 3 and 6. With these sets of candidates for each glimpse, we can now check for our second shortcut. If the total number of unique candidate locations over all glimpses is equal to the lower bound, then the numerosity is equal to the current lower bound. Images for which this shortcut are available receive a difficulty level of 1.

Based on the candidate sets, each glimpse is categorized as either unambiguous, if there is only one location in its set of candidates, or ambiguous otherwise. Without any further processing, the unambiguous glimpses tell us locations where there are definitely objects. We will update our estimate of the object locations $\hat{L}$ accordingly. The ambiguous glimpses can be further categorized as those that need to be resolved and those that do not. The ambiguity of an ambiguous glimpse does not need to be resolved if its set of candidate locations is contained within our current estimate $\hat{L}$. For example, in Figure 3, glimpses 1 and 3 are unambiguous, as indicated by the filled line that connects them to only one location. Glimpses 0 and 2 are both ambiguous, but only the ambiguity of glimpse 0 needs to be resolved because the candidates of glimpse 2 (indicated by the dashed lines) are locations already indicated by the unambiguous glimpses. If there are no ambiguous glimpses that need to be resolved, then we are done: our current estimate $\hat{L}$ but be equal to L and the numerosity must be $|\hat{L}|$. Images with no ambiguous glimpses to be resolved also receive a difficulty level of 1 because it required a single pass of computation through c_g to determine the numerosity.

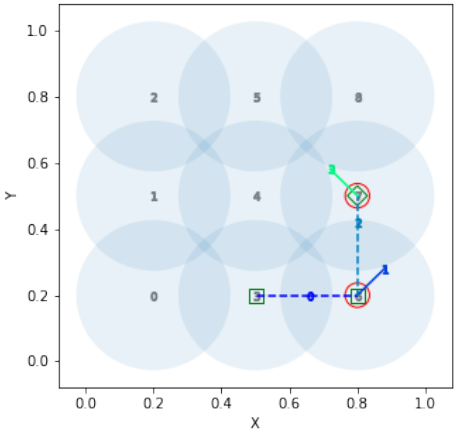

Figure A4: Knowledge state after processing the gaze coordinates. Filled lines indicate candidate locations for ambiguous glimpses. Filled lines indicate unambiguous glimpses. Red circles indicate locations where items are known to lie. G_1 and G_3 unambiguously reveal that there are items in locations 8 and 7 respectively. The candidate locations for G_2 are also 7 and 8, so G_2 is ambiguous but irrelevant. The only glimpse ambiguity that needs to be resolved is that of G_0.

Resolving ambiguity with the shape feature If there are ambiguous glimpses to be resolved, we use s_g to resolve that ambiguity. s_g is used to arbitrate among the hypotheses proposed by the initial processing of c_g. For example, recall in FigureA4, glimpse 0 is an ambiguous glimpse that needs to be resolved to determine the numerosity. Its candidate locations are 3 and 6. Location 6 has already been surmised to host an object based on information from the other glimpses. So the hypotheses are :

1. There is only an object in location 6. Location 3 is empty.
2. There are objects in both location 3 and location 6.

These hypotheses predict different patterns in s_g_0. If hypothesis 1 were true, s_g_0 would be all zeros except for one element which would be equal to the proximity of glimpse 0 to location 6. There are two ways that hypothesis 2 could be true. Either location 3 and 6 host the same type of shape, or they host different shapes. If two different shapes, s_g_0 will be all zeros except for two elements, whose values are equal to the glimpse's proximity to location 3 and 6. If the same shape, than s_g_0 will be all zeros except for one element whose value will be equal to the sum of the proximities of glimpse 0 to location 3 and location 6. These three predictions can be compared to the actual s_g_0 to determine which hypothesis is true. In general, the counter enumerates all hypotheses suggested by c_g_j and uses c_g_j to arbitrate among them. When the set of candidate locations is greater than 2 (can be max 4 in these simulations), the number of predicted patterns increases significantly since we need to enumerate all possible combinations of shapes.

After resolving an ambiguous glimpse, we update $\hat{L}$ and check if any of our previous stopping conditions are met. Note that resolving one ambiguous glimpse might render other ambiguities inconsequential as well, so we need to re-evaluate the set of ambiguous glimpses to be resolved. If any such glimpses remain, we pick another at random to be resolved and go back to the beginning of the loop.

The difficulty level will be incremented each time we resolve the ambiguity of an ambiguous glimpse. In these simulations, the number of glimpses is set to 4. So if all glimpses are ambiguous and need to be resolved separately, the difficulty level will be 5. If we were unable to completely resolve the ambiguity, which can happen if an image is one of a small number of edge cases that haven't been accounted for in the counter, the difficulty level is 6.

A glimpse cannot ever necessarily be uniquely assigned to an object—a glimpse is not 'of' an object per se—but each glimpse, because of how they were simulated, contains information about the set of objects in the image.