Zero-Shot Visual Numerical Reasoning in Dual-Stream Neural Networks

Jessica A. F. Thompson (jessica.thompson@psy.ox.ac.uk)
Department of Experimental Psychology, University of Oxford
United Kingdom

Hannah Sheahan (sheahan.hannah@gmail.com)
Department of Experimental Psychology, University of Oxford
United Kingdom

Christopher Summerfield (christopher.summerfield@psy.ox.ac.uk)
Department of Experimental Psychology, University of Oxford
United Kingdom

Abstract

Zero-shot numerical reasoning is challenging for modern computer vision systems but easy for humans. We present a dual-stream glimpsing recurrent neural network that combines gaze contents ("what") and gaze location ("where") to count the number of target items in a visual array, while ignoring distractors. The network successfully learns to count target items and generalizes to an out-of-distribution (OOD) test set 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. The model displays several neural response properties and patterns of behaviour that have previously been documented in primate visual enumeration. These results provide a proof-of-principle for a theory of the role of the parallel pathways of the primate visual system and posterior parietal cortex in visual relational reasoning.

Keywords: visual reasoning; dorsal stream; enactive cognition; zero-shot generalization; neural networks

Introduction

Human visual scene understanding involves processing the relations among objects. Two scenes composed of the same objects but in different relation can have very different meanings (e.g., consider a car in a box compared to a box in a car). Human viewers learn abstract concepts corresponding to relational properties and then can generalize these concepts zero-shot to new objects and contexts. Cardinality is a prime example of this ability. Once a child has learned the abstract concept of "threeness", she will forever be able to recognize groups of three objects, even novel objects in novel contexts, without any additional learning. This is not the case for modern neural network-based computer vision systems, which, while highly proficient at object recognition, struggle to generalize relational properties like cardinality (Zhang & Wu, 2020; Wu, Zhang, & Shu, 2019).

Recently, we showed that a recurrent dual-stream neural network, inspired by the role of the dorsal stream in primate vision, can perform zero-shot counting (Thompson, Sheahan, & Summerfield, 2022). Here, we focus on a more difficult numerical reasoning task—counting target objects while ignoring distractors. We find that a similar approach enables zero-shot generalization in this task. This work builds on previous research on glimpsing and dual-pathway neural networks (Larochelle & Hinton, 2010; Mnih, Heess, Graves, & Kavukcuoglu, 2014; Adeli, Ahn, & Zelinsky, 2022; Bakhtiari, Mineault, Lillicrap, Pack, & Richards, 2021; Mineault, Bakhtiari, Richards, & Pack, 2021)

Methods

We synthesized 48x42 pixel grayscale images consisting of 1–5 target letters and 0-2 distractor letters. The task is to report the number of targets, ignoring distractors. Training images consist of a subset of letters and luminance values not present in the test images, except for the letter A which is the distractor during training and at test. Letters lie on a 6x6 grid, are of constant size (5 pixels tall), and are never overlapping. All target shapes within one image are the same.

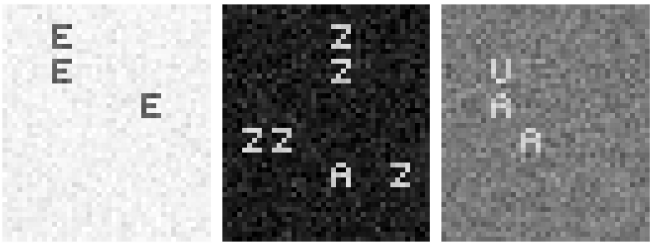

Figure 1: Example images consisting of 1–5 target items to be counted and 0–2 distractor items to be ignored.

Our dual-stream model, inspired by the parallel pathways of the primate visual system (Goodale & Milner, 1992) apprehends an image via a sequence of partial glimpses. Twelve glimpse locations are sampled from a saliency map consisting of a mixture of Gaussians with one Gaussian per item, subject to the constraint that every item in the image is glimpsed at least once. Both the glimpse locations (the gaze coordinates) and the glimpse contents (the 6x6 pixels centred at the glimpse location) are passed as input to the network. This design embodies the hypothesis that signals pertaining to a viewer's orientation to a scene, e.g. eye movements, contain

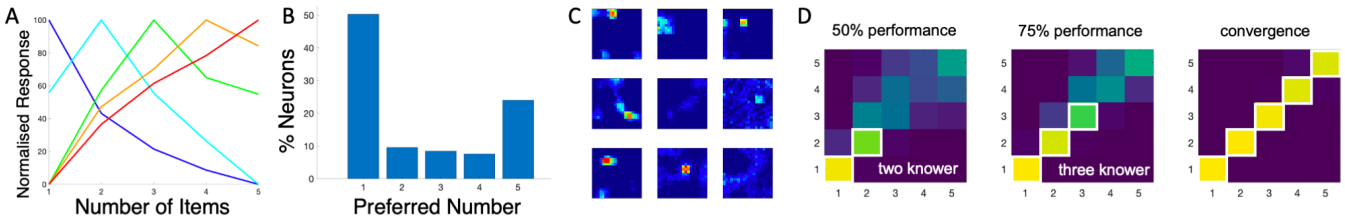

Figure 2: Dual-stream network replicates features of neural coding and development. A) grouping units by their preferred cardinality reveals log-normal number coding; B) more units are selective for the extremes of the number range; C) place coding-like spatial receptive fields; D) confusion matrices at increasing levels of proficiency.

content-invariant information about the scene's structure. The ventral module (which receives the glimpse contents as input) is pretrained to distinguish target and distractor shapes. A separate dataset including all letters and luminance values is used for ventral pretraining (although see the 'true OOD' condition described below). Ventral stream weights are not updated during training on the counting task. A recurrent module integrates the products of the ventral and dorsal modules over the 12 successive glimpses. The penultimate layer is trained via an auxiliary loss to produce a spatial map of the target items. This effectively decomposes the counting task into 1) localize the targets, then 2) count them.

Results

Our dual-stream model generalizes well, achieving 97.5% accuracy on the OOD test set ('dual stream' in Fig. 3). This zero-shot generalization behaviour is not observed in several parameter-matched control models which receive the entire image as input ('cnn' and 'mlp' in Fig. 3). Neither stream of the dual-stream model is sufficient to solve the task alone. Omitting either one of them renders the task unsolvable, reducing test accuracy to $\sim 45\%$ ('ablate ventral' and 'ablate dorsal' in Fig. 3). Test accuracy falls to $\sim 88\%$ without the ventral stream's auxiliary distractor recognition objective, allowing for ventral stream parameters to be updated with respect to the number and map objectives instead ('no ventral loss' in Fig. 3). The auxiliary map objective renders the counting task learnable—without it, training accuracy is at chance. This is unique to the Count Targets task. If instead the network is trained to count all items, including targets and distractors, the map objective is less crucial ('no map loss' and 'no map loss count all' in Fig. 3). As a stricter test of OOD generalization, the 'true OOD' condition uses a ventral stream that was pretrained only on those letters and luminances included in the training set for the Count Targets task. This reduces OOD test accuracy by ~ 10 percentage points, which is approximately equal to the generalization error on the ventral stream recognition task in this setting.

The dual-stream model replicates several neural and behvaioural phenomena associated with human and monkey enumeration. Analyzing the activity of the hidden layer of the recurrent module revealed several response properties associated with posterior parietal cortex (PPC): log-normal number coding (Fig 2A), an over-representation of units selective to the extremes of the number range (Fig 2B), and place-selective spatial receptive fields (Fig 2C), all of which have been documented in the PPC of rhesus monkeys (Nieder, Diester, & Tudusciuc, 2006; Viswanathan & Nieder, 2013, 2017). Inspection of the pattern of errors made at different levels of proficiency revealed that, like human learners (Sarnecka & Carey, 2008), the dual-stream model masters smaller numerosities first, gradually refining larger numerosities (Fig 2D).

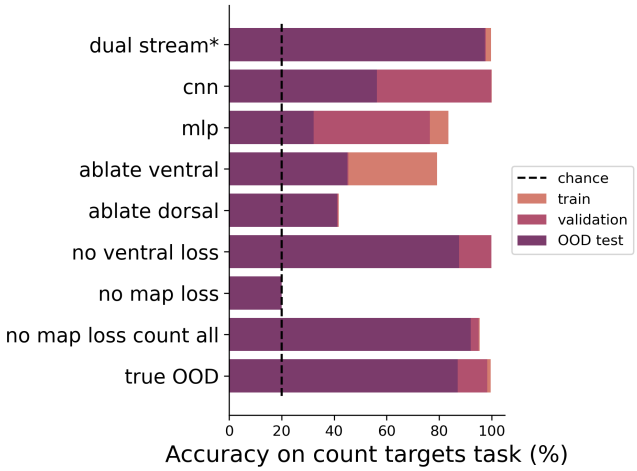

Figure 3: Model comparison of counting accuracy.

Discussion

Our characterization of the computational principles that support zero-shot numerical reasoning are consistent with converging theories of the role of posterior parietal cortex in visual reasoning (Summerfield, Luyckx, & Sheahan, 2020; O'Reilly, Ranganath, & Russin, 2021; Bottini & Doeller, 2020). According to this theory, efferent copies of motor or attention signals are received as an additional input to a visual reasoning system, enabling abstractions that are grounded in action, rather than purely in the sensory domain. The success of our model suggests that "attention" may be an enactive mechanism for relational inference, rather than (merely) a spatial prioritisation scheme.

Acknowledgments

This work was supported by generous funding from the European Research Council (ERC Consolidator award 725937) and Special Grant Agreement No. 945539 (Human Brain Project SGA). Thanks to David McCaffrey and Adam Harris for early discussions.

Diversity Statement

We sought to proactively consider choosing references that reflect the diversity of the field in thought, form of contribution, gender, race, ethnicity, and other factors. Using cleanBib (https://github.com/dalejn/cleanBib), we obtained the predicted gender of the first and last author of each reference by using databases that store the probability of a first name being carried by a woman (Dworkin et al., 2020; Zhou et al., 2020). By this measure (and excluding self-citations to the first and last authors of our current paper), our references contain 7.14% woman(first)/woman(last), 16.23% man/woman, 21.43% woman/man, and 55.2% man/man. Second, we obtained predicted racial/ethnic category of the first and last author of each reference by databases that store the probability of a first and last name being carried by an author of color (Ambekar, Ward, Mohammed, Male, & Skiena, 2009; Sood & Laohaprapanon, 2018). By this measure (and excluding self-citations), our references contain 17.05% author of color (first)/author of color(last), 3.35% white author/author of color, 33.14% author of color/white author, and 46.47% white author/white author.

References

Adeli, H., Ahn, S., & Zelinsky, G. (2022). A brain-inspired object-based attention network for multi-object recognition and visual reasoning. *bioRxiv*, 1–25.

Ambekar, A., Ward, C., Mohammed, J., Male, S., & Skiena, S. (2009). Name-ethnicity classification from open sources. In *Proceedings of the 15th acm sigkdd international conference on knowledge discovery and data mining* (pp. 49–58).

Bakhtiari, S., Mineault, P., Lillicrap, T., Pack, C. C., & Richards, B. A. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. In *Neurips* (p. 2021.06.18.448989).

Bottini, R., & Doeller, C. F. (2020). Knowledge Across Reference Frames: Cognitive Maps and Image Spaces. *Trends in Cognitive Sciences*, *24*(8), 606–619.

Dworkin, J. D., Linn, K. A., Teich, E. G., Zurn, P., Shinohara, R. T., & Bassett, D. S. (2020). The extent and drivers of gender imbalance in neuroscience reference lists. *bioRxiv*.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25.

Larochelle, H., & Hinton, G. (2010). Learning to combine foveal glimpses with a third-order Boltzmann machine. In *Neural information processing systems.*

Mineault, P. J., Bakhtiari, S., Richards, B. A., & Pack, C. C. (2021). Your head is there to move you around: Goal-driven models of the primate dorsal pathway. *bioRxiv*(NeurIPS), 2021.07.09.451701.

Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). Recurrent models of visual attention. *Advances in Neural Information Processing Systems*, 2204–2212.

Nieder, A., Diester, I., & Tudusciuc, O. (2006). Temporal and spatial enumeration processes in the primate parietal cortex. *Neuroforum*, *12*(4), 267–269.

O'Reilly, R. C., Ranganath, C., & Russin, J. L. (2021). The Structure of Systematicity in the Brain. *Current Directions in Psychological Science*, 1–7.

Sarnecka, B. W., & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition*, *108*(3), 662–674.

Sood, G., & Laohaprapanon, S. (2018). Predicting race and ethnicity from the sequence of characters in a name. *arXiv preprint arXiv:1805.02109*.

Summerfield, C., Luyckx, F., & Sheahan, H. (2020). Structure learning and the posterior parietal cortex. *Progress in Neurobiology*, *184*.

Thompson, J. A. F., Sheahan, H., & Summerfield, C. (2022). Learning to count visual objects by combining "what" and "where" in recurrent memory. In *Gaze meets ml neurips workshop - proceedings of machine learning research (pmlr)*.

Viswanathan, P., & Nieder, A. (2013). Neuronal correlates of a visual "sense of number" in primate parietal and prefrontal cortices. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(27), 11187–11192.

Viswanathan, P., & Nieder, A. (2017). Comparison of visual receptive fields in the dorsolateral prefrontal cortex and ventral intraparietal area in macaques. *European Journal of Neuroscience*, *46*(11), 2702–2712.

Wu, X., Zhang, X., & Shu, X. (2019). 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*, 1303–1310.

Zhang, X., & Wu, X. (2020). On numerosity of deep neural networks. *Advances in Neural Information Processing Systems*.

Zhou, D., Cornblath, E. J., Stiso, J., Teich, E. G., Dworkin, J. D., Blevins, A. S., & Bassett, D. S. (2020, February). *Gender diversity statement and code notebook v1.0.* Zenodo.