Neuron

Zero-shot counting with a dual-stream neural network model

Highlights

- We describe a dual-stream neural network model that displays zero-shot counting
- With ablations, we show how our dual-stream architecture supports this ability
- The model replicates several aspects of human counting behavior and development
- The learned representations mimic properties of neural codes for number and space

Authors

Jessica A.F. Thompson, Hannah Sheahan, Tsvetomira Dumbalska, Julian D. Sandbrink, Manuela Piazza, Christopher Summerfield

Correspondence

jessica.thompson@psy.ox.ac.uk (J.A.F.T.), christopher.summerfield@psy.ox.ac.uk (C.S.)

In brief

How does the brain represent the structure of a visual scene (the relations among items, e.g., the cardinality) independent of scene contents (the objects in the scene, e.g., item identity)? Thompson et al. propose a dual-stream neural network model based on the parallel pathways of the primate visual system.

Thompson et al., 2024, Neuron *112*, 1–12
December 18, 2024
https://doi.org/10.1016/j.neuron.2024.10.008

CellPress

Please cite this article in press as: Thompson et al., Zero-shot counting with a dual-stream neural network model, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.10.008

Neuron

CellPress
OPEN ACCESS

Article

Zero-shot counting with a dual-stream neural network model

Jessica A.F. Thompson,[1,3,*] Hannah Sheahan,[1] Tsvetomira Dumbalska,[1] Julian D. Sandbrink,[1] Manuela Piazza,[2] and Christopher Summerfield[1,*]

[1]Department of Experimental Psychology, University of Oxford, Oxford OX2 6GG, UK
[2]University of Trento, Department of Psychology and Cognitive Science, Trento 38068, Italy
[3]Lead contact
*Correspondence: jessica.thompson@psy.ox.ac.uk (J.A.F.T.), christopher.summerfield@psy.ox.ac.uk (C.S.)
https://doi.org/10.1016/j.neuron.2024.10.008

SUMMARY

To understand a visual scene, observers need to both recognize objects and encode relational structure. For example, a scene comprising three apples requires the observer to encode concepts of "apple" and "three." In the primate brain, these functions rely on dual (ventral and dorsal) processing streams. Object recognition in primates has been successfully modeled with deep neural networks, but how scene structure (including numerosity) is encoded remains poorly understood. Here, we built a deep learning model, based on the dual-stream architecture of the primate brain, which is able to count items "zero-shot"—even if the objects themselves are unfamiliar. Our dual-stream network forms spatial response fields and lognormal number codes that resemble those observed in the macaque posterior parietal cortex. The dual-stream network also makes successful predictions about human counting behavior. Our results provide evidence for an enactive theory of the role of the posterior parietal cortex in visual scene understanding.

INTRODUCTION

The meaning of a visual scene depends on both its contents and its structure. The contents of a scene are the objects it contains. For example, in each panel of Figure 1A, there are two salient objects: cats and bowls. The structure of a scene defines how the objects relate to each other. The meaning of each panel in Figure 1A depends on whether there are more cats than bowls or vice versa, and whether the arrangement of objects is orderly or disorderly. The importance of object relations for understanding scene structure has been appreciated for at least a century, since the first investigations of Gestalt psychology.[1]

Over recent years, we have learned a great deal about the computations that underlie the recognition of lone objects presented briefly at the fovea. Lesion and recording studies imply that object recognition relies on ventral visual regions of the primate brain.[3,4] The mapping of naturalistic images to semantic labels can be modeled as a feedforward cascade through successive processing layers of a neural network.[5–9] Deep convolutional networks trained with gradient descent to label images develop neural population codes that roughly match those observed in electrophysiology and neuroimaging studies of the primate ventral stream.[10–13] This success with modeling perception of scene *contents* notwithstanding, computational models of how the relational *structure* of a scene is processed remain much less mature.[2,14,15]

One major challenge is that humans can immediately apprehend many relational aspects of a visual scene even if the objects it contains are wholly novel. For example, you may not recognize the objects in Figure 1B but have no difficulty reporting that there are three of them. Here, we call this phenomenon "zero-shot counting." We use "zero-shot counting" to refer to the ability of an animal or artificial system to count items in visual arrays whose perceptual features differ significantly from those of the scenes on which the agent learned to count. This is an instance of out-of-distribution (OOD) generalization.

This ready ability to apprehend structure (object relations) without being able to recognize contents (object identity) is puzzling in the context of deep learning models, which often fail dramatically when probed about structural aspects of a novel scene. For example, neural networks struggle to identify whether two previously unseen objects are the same or different.[16] In the case of counting, supervised learning of numerosity is severely disrupted when the objects being counted lie outside the training distribution.[17,18] Even very large generative models are prone to make structural errors in scene composition (including numerosity) when mapping text to images.[19,20] This implies that understanding scene structure relies on computational processes that are not currently included in the canonical deep learning framework for modeling object recognition.

In humans, correctly inferring relations among objects in a scene depends on the integrity of dorsal stream structures, including the posterior parietal cortex (PPC). For example, patients with bilateral damage to the PPC often have difficulty

Please cite this article in press as: Thompson et al., Zero-shot counting with a dual-stream neural network model, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.10.008

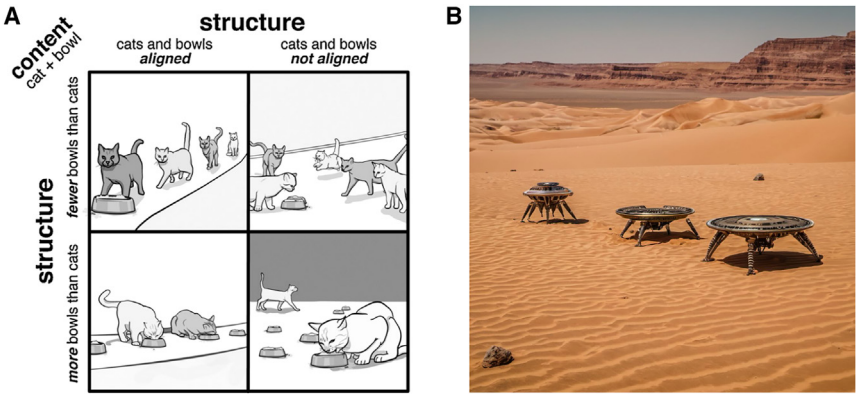

Figure 1. Cognitive phenomenon of interest

(A) The meaning of a visual scene depends on both its content and structure. Image by Hannah Sheahan,[2] reprinted with permission.

(B) Humans who master the cardinal principle will have no difficulty counting completely novel objects in novel contexts without needing any additional training examples—they generalize numerosity zero-shot. Image generated with https://www.img2go.com/.

counting, comparing, or localizing objects in a visual array. One possibility is that biological brains evolved a visual system that factorizes the processing of scene contents and structure into respective ventral and dorsal pathways.[2] In this theory, the ventral stream supports object recognition (a representation of "what") and the dorsal stream learns to code explicitly for regions of space (a representation of "where"). Explicit, factorized codes for what and where would allow the brain to make inferences about spatial structure that are not tied to existing object representations.

In the current work, we describe a neural network model that implements this idea with a recurrent dual-stream architecture. The key idea is that a neural network can learn explicit representations of space—similar to those found in the primate dorsal stream—by recycling its outputs (here, signals controlling glimpses or simulated saccadic eye movements) as inputs for representation learning. This emphasizes that agents can learn about the structure of space by taking actions and, in doing so, this allows them to make inferences about the structure of a scene, even when unfamiliar with its contents. We show that this network can solve the "zero-shot counting" problem and that, in doing so, it develops neural representations that closely resemble those observed in the PPC of the nonhuman primate. The network also successfully predicts new behavioral results observed in human participants performing an eye-tracking task that involves counting objects among distracters.

RESULTS

We call the problem we set out to solve zero-shot counting. It is operationalized as follows. The observer is asked to classify the number (1–5) of target items in a two-dimensional (2D) grayscale image (Figure 2A), potentially in the presence of up to two prespecified distracter items. Both targets and distracters are alphanumeric characters embedded in a pixelated background. Foreground and background luminance values are sampled from Gaussian distributions whose means differ by at least 30% (Figure 2A). To ensure that the task cannot be partially solved by counting the number of unique letters glimpsed, all target items within the array are the same letter. There is only one class of distracter (the letter A). During training, we sample targets from set T_{train} (B,C,D,E). At test, with no further supervision signals provided, we evaluate the network counting performance for targets sampled from disjoint set T_{test} (F,G,H,J). We also allow the distribution of mean luminance values to potentially vary between training and test (giving us l_{train} and l_{test}).

Zero-shot counting performance

In Figure 2B we show the performance of a standard convolutional neural network (CNN) on this zero-shot counting task. We begin with the simplest case in which no distracters are present ("simple counting"). While training on a set of images defined by $\{T_{train}, l_{train}\}$, we evaluate on new images drawn from the same distribution (validation) as well those drawn from a new distribution of luminance values (OOD luminance), new letters (OOD shape), or both (OOD both). Stimuli parameters for each dataset are listed in Table 1. The CNN successfully learns to count the items in the training data and can generalize this to the validation set (accuracy = 99.9% ± 0.01%, chance = 20%) but not to the OOD conditions T_{test} (OOD luminance, accuracy = 81.6% ± 17.7%; OOD shape, accuracy = 72.1% ± 6.7%; OOD both, accuracy = 63.0% ± 13.5%). All mean accuracy and standard deviation values are calculated over 20 different random seeds. Accuracy on each OOD test set was less than that for the validation set (one-sided Wilcoxon signed-rank tests, all $w = 210$, $p < 0.001$; all p values Bonferroni corrected).

In the case where zero, one, or two distracters are present ("ignore distracters"), the CNN again performs well on the validation set (mean accuracy = 99.9% ± 0.01%) but its accuracy is dramatically reduced by OOD stimuli, especially in the OOD shape condition (23.5% ± 0.10%) and OOD both condition (25.5% ± 1.8%; accuracy in OOD luminance condition is 81.7% ± 8.5%). Note that chance is 20% for these tasks. The CNN thus fails at zero-shot counting both with and without distracters. The finding that the CNN is perturbed by changes to irrelevant features of the image, such as luminance, might imply that CNNs solve the mapping problem by representing textural features.[21,22] However (without further constraints), CNNs do not naturally individuate objects as humans do when computing object relations and thus struggle during counting of new objects.

The architecture of our proposed dual-stream recurrent neural network (RNN) model (shown in Figure 3B) is inspired by the structure of the primate visual system, highlighted in Figure 3A. First, our network samples the image in a quasi-naturalistic way. Unlike the standard CNN, which receives the whole image at once as input, our network views each image through a sequence of

Please cite this article in press as: Thompson et al., Zero-shot counting with a dual-stream neural network model, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.10.008

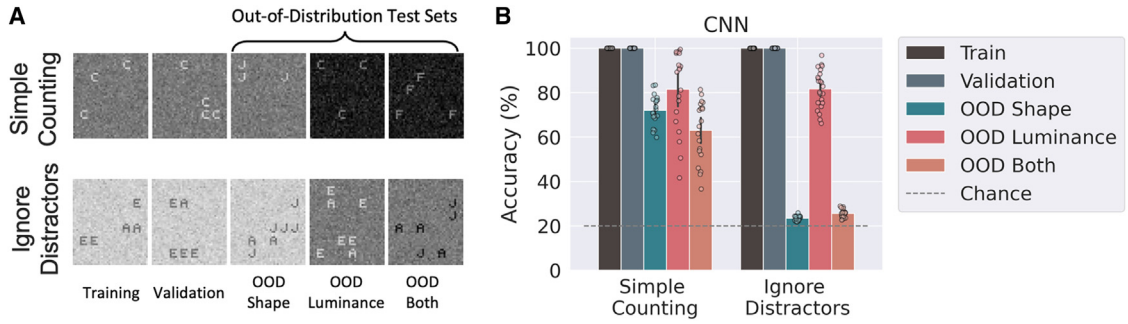

Figure 2. Task and baseline performance

(A) Example images for each task and dataset. For "simple counting," there are no distracter items (no letters A). The validation set is independently sampled from the same distribution as the training set. The OOD test sets contain letters and/or mean luminances that were not present in the training set.

(B) Number classification accuracy for a convolutional neural network. Each dot represents one model run with a different random seed. Error bars indicate bootstrapped 95% confidence intervals. See Table 1 for stimuli parameters and Table S3 for CNN model parameters.

spatially localized glimpses according to a biologically realistic gaze policy. It processes each glimpse with higher resolution at the locus of fixation, mimicking the primate fovea.[23,24] In the primate, glimpse contents are fed forward from the thalamus and V1 to the ventral stream structures such as the V4 and temporal cortex areas, including the inferior temporal cortex (IT/TE). Like others, we model this ventral stream processing with a convolutional architecture (Figure 3B, green box) (in the "ignore distracters" condition, we pretrain this module to distinguish targets and distracters). The output of the primate ventral stream flows to the higher association cortex, such as the PPC, which (along with prefrontal cortex) is thought to integrate information across a sequence of glances in visual short-term memory.[25] To mimic this, in our model, outputs from the "ventral" convolutional module are fed forward to a recurrent module that we equate with the PPC. Recurrent computation allows information about number to be combined across glimpses (Figure 3B, yellow box), and the proposed connectivity is consistent with known pathways from the IT to the PPC.[26–28]

The key feature of our model, however, is that glimpse contents (what) are processed in parallel with glimpse position (where). We implement this dual-streams principle in the simplest way possible: on each saccade, we simply pass the (x, y) location of the glimpse to the recurrent module. The two input streams are merged in a joint embedding layer, which then feeds into an RNN submodule. In the primate brain, we know that the superior colliculus (SC) encodes a topographic map of salient regions of visual space and computes a gaze vector, which is responsible for driving saccadic eye movements.[29] We also know that the SC is reciprocally connected with the PPC via the pulvinar,[30–32] providing a putative pathway for the recurrent module to receive information about the current position of the eyes. Two further layers successively process outputs from the RNN submodule. The penultimate layer of the network is trained with an auxiliary loss to produce a spatial map of the location of target items in the image (we call this the "map layer"). From the map layer, a linear read-out classifies the numerosity in the scene (Figure 3B).

Consistent with a previously described theory,[2] we reasoned that this architecture would be able solve the zero-shot counting task because during training it would learn representations that explicitly combine information about visual contents (what; glimpse pixels from the ventral stream) and structure (where; glimpse position from the dorsal stream). We predicted that this would allow the network to generalize across scene structure (numerosity) even where scene contents (objects) were entirely novel. We also previously proposed that this architecture would help explain the coding properties of neurons in the PPC and closely interconnected regions.[2]

In Figure 3C, it can be seen that the dual-stream RNN is indeed able to solve the zero-shot counting problem. On the "simple counting" validation set, its performance is comparable with the CNN (99.0% ± 0.1%). However, unlike the CNN, it maintains this performance across OOD shapes (98.1% ± 0.2%), OOD

Table 1. Training and test set parameters

Dataset	Target shapes	Distracter shape	Mean luminances	# Images
Training	B, C, D, E	A	0.1, 0.4, 0.7	100,000
Validation	B, C, D, E	A	0.1, 0.4, 0.7	5,000
OOD shape	F, G, H, J	A	0.1, 0.4, 0.7	5,000
OOD luminance	B, C, D, E	A	0.3, 0.6, 0.9	5,000
OOD both	F, G, H, J	A	0.3, 0.6, 0.9	5,000

Dataset parameters are the same for both "simple counting" and "ignore distracters" except that no distracters are present in the "simple counting" images.

Please cite this article in press as: Thompson et al., Zero-shot counting with a dual-stream neural network model, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.10.008

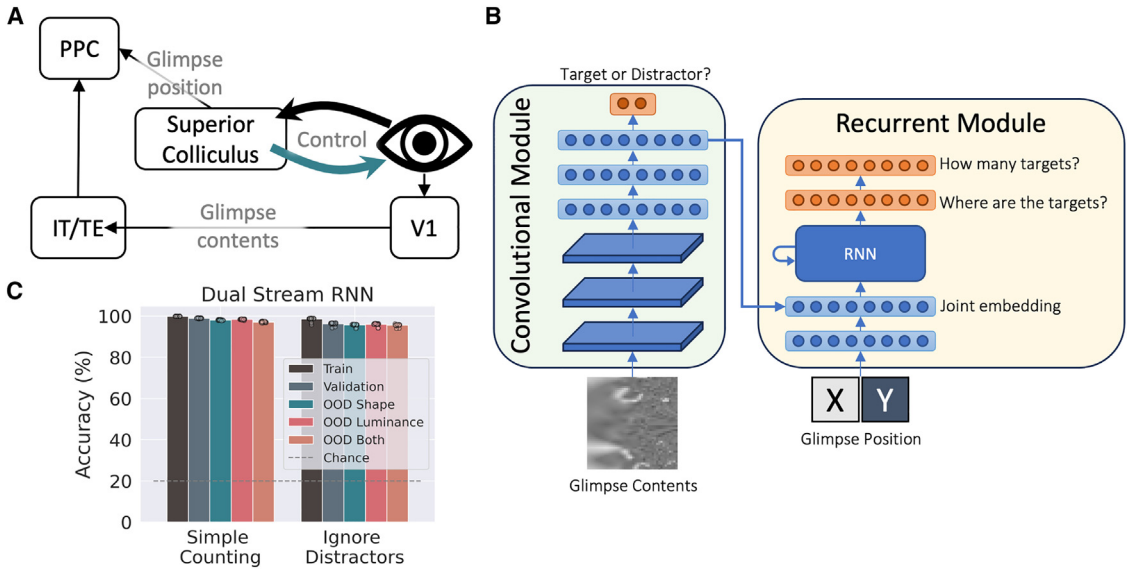

Figure 3. Dual-stream RNN model

(A) Schematic of the relevant components of the primate visual system. Efferent copies of motor instructions for intended eye movements propagate PPC via the superior colliculus. PPC integrates glimpse contents from the ventral stream and glimpse positions from superior colliculus.

(B) Dual-stream RNN architecture. A convolutional module receives the foveated glimpse contents and is pretrained to distinguish target and distracter shapes. In parallel, a recurrent module receives the glimpse positions via a separate input stream. The recurrent module integrates the two streams over successive glimpses to produce a map of the spatial arrangement of target items in the array from which the number of items is read out. Orange layers indicate where losses are calculated.

(C) Number classification accuracy for a dual-stream RNN. Each dot represents one model run with a different random seed. Error bars indicate bootstrapped 95% confidence intervals. See Table S1 and Table S2 for detailed model parameters. Also see Figure S1 for additional comparisons between the dual-stream RNN and CNN baseline.

luminances (98.4% ± 0.2%), and OOD both (97.1% ± 0.3%) conditions. It performs comparably in the "ignore distracters" condition (OOD shape 95.6% ± 0.5%; OOD luminance 96.0% ± 0.6%; OOD both 95.5% ± 0.6%). For the dual-stream RNN, the mean difference between validation performance and OOD both performance was less than 1 percentage point (0.8% ± 0.3%) for "simple counting" and similar for "ignore distracters" (1.9% ± 0.3%). By contrast, for the CNN, the mean difference between validation and OOD both performance was 37% ± 13.2% for "simple counting" and 74.5% ± 1.7% for "ignore distracters." Thus, the dual-stream RNN is able to solve the zero-shot counting task where the CNN is not. In Figure S1, we show that this pattern holds even when using larger, more perceptually diverse images.

Ablations and controls

Next, we conducted control analyses that pinpoint those neural or computational features of our architecture that are critical for its success (Figure 4). First, we carried out virtual lesion studies to examine the causal role of ventral inputs (glimpse contents) and dorsal inputs (glimpse positions) on network performance. We performed lesions by removing either glimpse contents or glimpse position inputs during both training and test. In the case of "simple counting," dorsal stream lesions ("ablate position" in Figure 4) were more detrimental (reducing performance on OOD both to 48.0% ± 2.5%) than ventral stream lesions ("ablate contents" in Figure 4) (80.4% ± 0.3% on OOD both). This is consistent with the finding from neuropsychological studies that PPC lesions lead to counting deficits, whereas temporal lobe lesions have a much milder impact.[33–35] By contrast, in the "ignore distracters" task, lesioning either the dorsal or ventral stream had a dramatic effect on performance. On the OOD both generalization condition, classification accuracy was reduced to 42.1% ± 0.4% by ventral stream lesions and to 52.7 ± 0.3 by dorsal stream lesions (more than double the error rate observed in "simple counting"). Although we are not aware of neuropsychological data that directly support this finding, there is good evidence that ventral stream lesions impair configural learning when object arrays become more complex.[36,37]

Why does counting the number of objects in a scene (that is apprehended through a series of glimpses) require both "what" and "where" information? Intuitively, glimpse contents alone are often insufficient, especially when all items in the array are identical. For example, if the network glimpses three items in succession, without auxiliary glimpse position inputs, it is unclear whether it has glimpsed three unique items (*i*, *j*, *k*) or glimpsed two before returning to the first item (*i*, *j*, *i*). However, gaze position alone is insufficient for counting because saccades are not exclusively directed precisely to the center of a single item but often fall in an intermediate zone between two or more items, allowing the network to apprehend them in a single glimpse. Ventral stream lesions are especially detrimental in the "ignore distracters" task because glimpse

Please cite this article in press as: Thompson et al., Zero-shot counting with a dual-stream neural network model, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.10.008

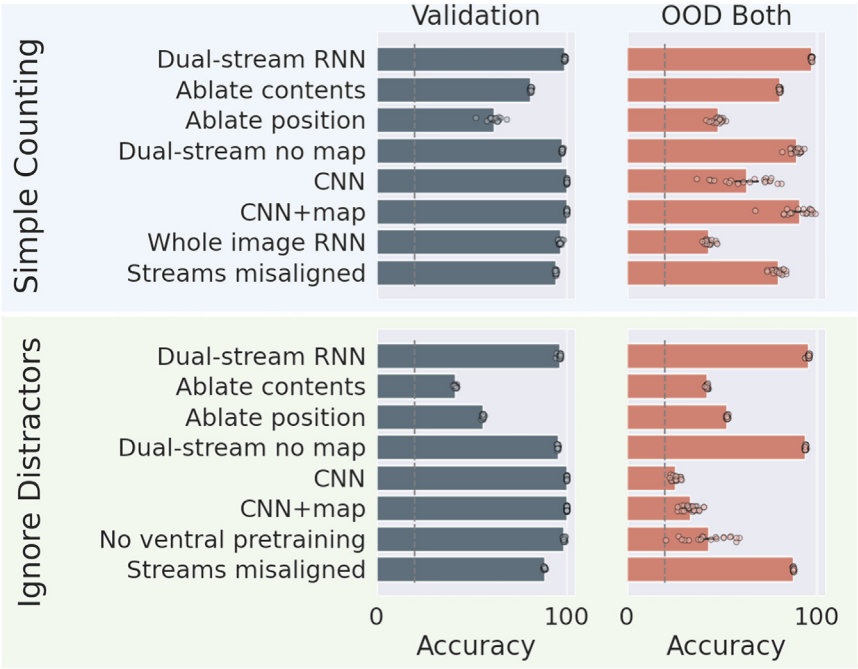

Figure 4. Ablations and control models

Accuracy on the validation set (left) and OOD both test set (right) for various control models and ablations. Each row corresponds to a particular training recipe or configuration of either the dual-stream RNN or the CNN baseline. Each dot is one model run with a different random seed. Error bars show bootstrapped 95% confidence intervals. When no error bar is present, the 95% CI is too small to be visible. Data are shown separately for the "simple counting" (upper) and "ignore distracters" (lower) conditions. See Figures S2 and S3 for additional controls.

contents are needed to signal whether an item is a target or distracter.

To study exactly how the network needs to combine what and where information in the task, we built a symbolic counter model that used a sequence of rules to classify the number of items given a set of glimpses (positions and contents) in a simplified version of the task. The counter worked by first attempting to infer the location of items from the glimpse positions alone and then querying the glimpse contents only to resolve any remaining ambiguity. Thus, the number of queries to the glimpse contents provided an "integration score" for each glimpsed image, indicating the degree to which the two input streams need to be integrated to solve the task. For example, if the set of glimpses contained only two unique positions in opposite corners of the image, it would be relatively clear from the glimpse positions alone that there are 2 items in the image. This would receive a low integration score. If, on the other hand, there are several glimpses clustered in a region that could reasonably contain 1, 2, or 3 items, the counter would need to inspect the contents of those glimpses to know for sure, prompting a higher integration score. This counter allows us to pinpoint the impact of both ventral and dorsal lesions on the counting process (Figure S2). When ablating the glimpse contents input, generalization performance scales inversely with integration score. The glimpsed images that the model struggles with in this one-stream setting are exactly those that the symbolic model identified as requiring an integration of both streams due to an ambiguity in the glimpse positions regarding item location.

Next, to evaluate the contribution of the various neuro-inspired architectural features of the dual-stream RNN, we compared several control models in which we disrupt some feature of the dual-stream RNN or add components to the CNN baseline. Our dual-stream RNN differs architecturally from our baseline model (a vanilla CNN) in that it comes equipped with recurrent memory. We first confirmed that recurrence alone was insufficient to solve the zero-shot counting problem. To test this, we created a one-stream version of the dual-stream RNN in which (like the CNN) the full image was input on each successive glimpse ("whole image RNN" in Figure 4). Like the CNN, this control network was also significantly impaired on all three OOD conditions ("simple counting": OOD luminance 54.6% ± 4.1%; OOD shape 48.4% ± 1.7%; OOD both 42.9% ± 2.2%). Omitting the ventral stream objective, instead allowing all network parameters to be updated end-to-end with respect to the number and map objectives ("no ventral pretraining" in Figure 4), greatly reduced its generalization performance on the "ignore distracters" task (OOD both 42.9% ± 11.3%). Adding the auxiliary map objective to the CNN ("CNN + map" in Figure 4) improved its generalization performance during "simple counting" (OOD both 90.7% ± 7.8%) but was of little help in "ignore distracters" (OOD both 33.6% ± 3.1%). Removing the auxiliary map loss from the dual-stream RNN ("dual-stream no map" in Figure 4) had a much more modest impact on performance (OOD both: "simple counting" 85.3% ± 16.0%, "ignore distracters" 93.9% ± 0.2%). See Figure S3 for a more detailed comparison of the impact of the auxiliary map loss. Generalization performance was also negatively impacted by shuffling the order of the glimpse positions relative to the glimpse contents ("streams misaligned" in Figure 4) ("simple counting": OOD both 79.4% ± 2.9%; "ignore distracters": OOD both 87.6% ± 0.3%). All mean accuracies and standard deviations calculated over 20 random seeds.

Together, these analyses show that both the dorsal and the ventral stream are necessary for solving the zero-shot counting task. We were unable to identify a trivial computational feature that can be added to a standard network to account for this success.

Comparisons with human behavior

Next, we studied the behavior of the dual-stream RNN across the trajectory of learning. As children learn to count objects in visual scenes, they often pass through discrete phases in which they

Please cite this article in press as: Thompson et al., Zero-shot counting with a dual-stream neural network model, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.10.008

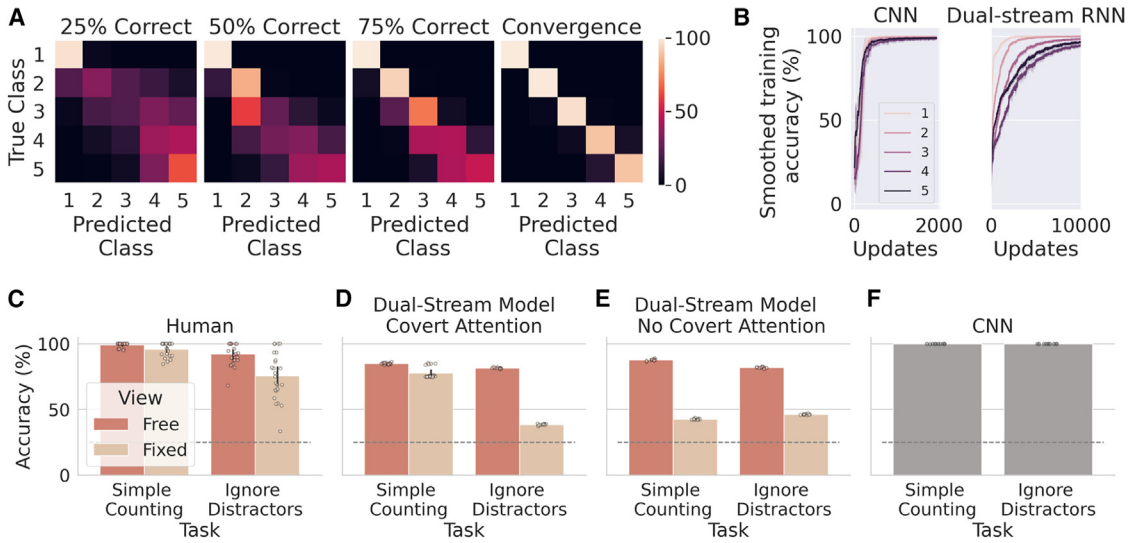

Figure 5. Comparisons with human behavior

(A) Confusion on the validation set ("ignore distracters") for the dual-stream RNN at checkpoints throughout learning. Rows specify the actual number of items and columns specify the predicted number of items, according to the dual-stream RNN. The color indicates the proportion of images of number class y that were predicted to be of number class x (light = high, dark = low), averaged over 20 repetitions with different random seed. Rows sum to 100%.

(B) Smoothed mini-batch training accuracy per number class (1–5; for "simple counting"). This is the accuracy encountered before each weight update, which permits inspection of the learning curves at a fine temporal resolution. Time courses were smoothed with a rolling average of 25 updates. Error bars indicate standard deviation over five repetitions with different random seed.

(C) Human performance on the counting tasks. Each dot corresponds to a particular participant's performance in one of four conditions. Dashed line indicates chance. Error bars are bootstrapped 95% confidence intervals.

(D and E) Validation accuracy of the dual-stream RNN with covert attention (D) and without (E). Each dot is one model run with a different random seed. Error bars are bootstrapped 95% confidence intervals.

(F) Validation accuracy of the CNN. Here, there is only one bar per task because there is no way of simulating "free" or "fixed" gaze conditions for the CNN. Each dot represents one model run from different random seed. See Figure S4 for further comparison with human behavior.

can enumerate two, three, or four items before grasping the general principle of cardinality. These phases are sequenced such that children are "two knowers," "three knowers," and "four knowers" before graduating to become "cardinal principle knowers."[38] We tested for this pattern in the behavior of the dual-stream RNN over the course of training. In Figure 5A, we show matrices that reveal the pattern of confusions (on the validation set) it makes for items with different numerosities for checkpoints lying at 25%, 50%, and 75% accuracy and at the end of the training run. In Figure 5B, we plot training curves for arrays with a ground truth number of 1–5 items. It can be seen the dual-stream RNN, but not a CNN, learns to accurately classify arrays with a smaller number of items first.

When asked to count the number of objects in a visual array, humans are prone to biases that depend on the regularity, spacing, and similarity of the items being counted. For example, humans tend to overestimate the number of items in an array when they are oriented more similarly to each other than when they are different.[39] We tested whether our dual-stream RNN displayed the same bias by training it on a mixture of homogeneous and heterogeneous arrays and measuring its bias to over- or under-estimate numerosity at test. We found that, like humans, the model exhibited a bias to overestimate the tally when items were more homogeneous (fewer unique items) (Figure S4).

Our theory seemingly makes a counter-intuitive prediction: that humans' ability to count items in a visual array depends on our capacity to move our eyes. We know that eye movements and visual counting are linked. For example, patients with bilateral damage to the PPC suffer from Balint's Syndrome, whose symptoms combine optic ataxia (disrupted saccadic eye movements) with simultanagnosia (an inability to perceive more than one object at a time).[40] Nevertheless, people can perceive small numerosities in a single glance (such as when you read a number five off a die), an ability that is known as "subitizing."[41] This seems to present a challenge for our theory. We thus conducted an eye-tracking experiment involving human participants to ask whether our network was able to predict patterns of human counting performance under free and fixed gaze.

We asked human participants (n = 24) to perform a visual counting task while we tracked their gaze position on the screen. We crossed task ("simple counting" vs. "ignore distracters") with gaze (free vs. fixed) in a 2 × 2 within-subjects design. Stimuli contained 3–6 target items and, in the "ignore distracters" task, 1–3 distracter items. In free gaze blocks, participants could move their eyes as they wished, whereas in fixed gaze blocks, they were obliged to maintain their eyes within 100 pixels of central fixation during counting or else the trial was aborted (and repeated at the end of the block). We found that, in humans, fixing the gaze impaired performance to a much greater degree in the "ignore distracters" task than in "simple counting" (Figure 5C). This observation was qualified by a two-way ANOVA on accuracy, which

Please cite this article in press as: Thompson et al., Zero-shot counting with a dual-stream neural network model, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.10.008

CellPress
OPEN ACCESS

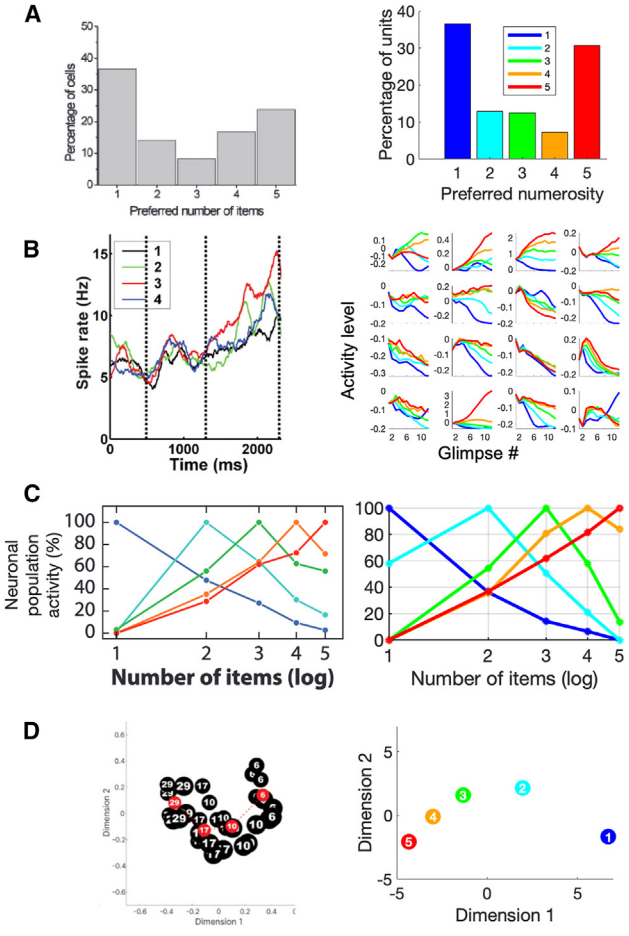

Figure 6. Neural coding

The dual-stream RNN (right) mimics canonical signatures in neural population codes for number recorded in primate PPC (left).

(A) Left: frequency distributions of the preferred numerosities for macaque PPC from Nieder and Miller,[47] reprinted with permission (Copyright [2004] National Academy of Sciences, USA). Right: frequency distributions of preferred numerosities among the recurrent units of the dual-stream RNN.

(B) Left: spike density histogram for an IPS neuron from Nieder et al.,[45] reprinted with permission. Right: mean responses to each number class as a function of glimpse number for a random sample of units in the recurrent layer of the dual-stream RNN.

(C) Left: Gaussian tuning curves displaying lognormal number coding from Nieder and Dehaene,[42] reprinted with permission. Right: average tuning of units in the recurrent layer of the dual-stream RNN.

(D) Left: MDS of BOLD (blood-oxygen-level-dependent) activity from human IPS during dot counting from Karami et al.[46] The black circles represent each stimulus labeled according to their numerosity (6, 10, 17, and 29) and scaled in size to reflect the total area of the dots. The red circles indicate the average coordinates of each number (CC-BY 4.0 International license). Right: MDS applied on the population activity of recurrent units in the dual-stream RNN. See Figures S5 and S6 for tuning curves and Figure S7 for the dimensionality analysis.

revealed a significant interaction between task and gaze ($F(1, 92) = 10.23$, $p < 0.01$); there was a significant reduction in accuracy from free to fixed conditions on the "ignore distracters" task (mean diff = 16.6, Tukey's honestly significant difference (HSD) adjusted $p < 0.001$, family-wise error rate [FWER] = 0.05) but not the "simple counting" task (mean diff = 3.3, Tukey HSD adjusted $p = 0.6904$, FWER = 0.05).

We simulated these data using our dual-stream RNN. On fixed trials, the network repeatedly glimpsed the center of the screen, whereas free trials unfolded exactly as described above (for this simulation, we trained the network on an equal mixture of free- and fixed-viewing trials so that neither gaze condition was out of distribution during test). We considered two settings: a "covert attention" setting where the network continues to receive putative gaze location information, even when it is forced to fixate centrally, and a "no covert attention" setting, where no gaze location information is offered on fixed trials. With covert attention, the network recreated the exact performance pattern observed in the human data—that performance on "ignore distracters" was affected by enforcing central fixation to a much greater degree than on "simple counting" (Mann-Whitney $U = 120$, $p < 0.001$; Figure 5D). This was not the case if covert attention was removed (Mann-Whitney $U = 0$, $p = 0.99$). In this case, enforcing fixation affected both tasks almost equally (Figure 5E). By contrast, the baseline CNN, which has no fovea, achieves 100% accuracy on the in-distribution validation set for both tasks and, as such, is unable to account for the pattern of errors that humans make in these conditions (Figure 5F).

Comparisons with neural codes for number

A natural next question is whether principles of neural coding in our network match those observed in the primate brain. We focused on responses in the recurrent layer, which we equate with the primate PPC. There are at least four canonical signatures for numerosity that are detectable in neural population codes recorded in the macaque and human PPC (many of which are replicated in prefrontal regions)[42,43] (Figure 6, left). First, neurons are tuned to number, and more neurons prefer the smallest and largest numbers in a discrimination set compared with those in between[44] (Figure 6A). Second, when making numerosity judgments, some number-selective units in the intraparietal sulcus (IPS) display firing rates that ramp up over time, often with the steepest slopes for the largest numbers[45] (Figure 6B). Third, neurons tend to code for number with approximately bell-shaped (Gaussian) tuning curves, and tuning width grows with number[42] (log-normal number coding; Figure 6C). Finally, at the population level, neural codes are low dimensional, tracing out a neural "number line" that becomes visible when each number is expressed as a point in a space with just a few axes, derived with dimensionality reduction techniques[46] (Figure 6D).

We found that even without any hyperparameter tuning, the dual-stream RNN naturally recreates each of these neural coding motifs (Figure 6, right). We show example tuning curves for numbers (1–5) in Figure S5. When we plot the frequency distribution of preferred selectivity, we can see that more cells prefer the extremes of the tested number range, a phenomenon that is also observed in the PPC[47] (Figure 6A). Individual units develop preferences for different numbers, which tend to ramp up or down over successive saccades (Figure 6B). This is similar to the pattern observed in the macaque PPC when monkeys make numerosity judgments about arrays of dots.[42,44,48] A salient feature of number coding in the PPC is that when the tuning curves of

Please cite this article in press as: Thompson et al., Zero-shot counting with a dual-stream neural network model, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.10.008

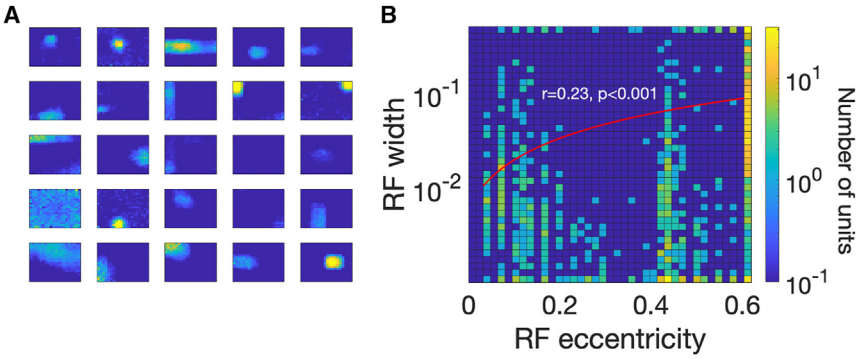

Figure 7. Spatial response fields

(A) Example spatial response fields (RFs) of units in the recurrent layer of the dual-stream RNN. Warmer colors signal higher mean activation.

(B) 2D histogram of the width (as fraction of image width) and eccentricity (as Euclidean distance from the image center [0.5, 0.5]) of the mean of isotropic Gaussian of best fit for each unit's RF. Color indicates the number of units. Red line shows line of best fit ($r = 0.23$, $p < 0.001$). RF width and number of units are presented on a log scale.

cells are sorted and averaged, the coding is more precise for lower numbers, resulting in a characteristic "log-normal" code.[42] When we perform the same analysis, we see an identical phenomenon (Figure 6C). Indeed, our average tuning curves were better fit by a model in which tuning curves were Gaussian in the space of $log(n)$ rather than n itself (F-ratio test, $F(15, 15) = 3.16$, $p < 0.05$) (Figure S6). Moreover, it is well known from macroscopic recordings that the neural similarity of the population response to number is well described by a single dimension, known as the number line.[46,49] We show the multidimensional scaling (MDS) projection of the population activity in 2D in Figure 6D, which reveals a curved number line. To test the dimensionality of the population data, we split trials into two halves and attempted to systematically reconstruct one half from dimensionality-reduced versions of the other. Figure S7 shows the mean variance explained on the held-out half, as a function of the dimensionality of the half used for training. The best reconstruction was obtained when just three dimensions remained. Therefore, we infer from the MDS projection and the dimensionality analysis that the dual-stream RNN has learned a low-dimensional representation with a mental number line in the first dimension.

Next, we examined the spatial representations that formed in the network. Cells in the PPC exhibit spatial response fields (RFs) in which firing rates are elevated in a temporal window surrounding a saccade made to a particular location.[50,51] In Figure 7A, we show example spatial RFs for neurons in the recurrent layer of the dual-stream RNN, many of which display the kind of spatial selectivity observed in PPC RFs. The model RFs exhibit a well-known property of PPC RFs—neurons tuned to more eccentric locations have broader fields.[52] When we fit isotropic Gaussians to each observed spatial RF, we observed a positive correlation between the eccentricity of the best-fitting mean (Euclidean distance to image center) and the best-fitting standard deviation (or RF width) ($r = 0.23$, $p < 0.001$) (Figure 7B). This correlation remains positive and significant when we exclude the units with RF eccentricity greater than 0.6 ($r = 0.10$, $p < 0.01$).

DISCUSSION

The findings described here support an "enactive" view of cognition in which motor signals (here, eye movements) are inputs to, as well as outputs from, the computations that determine how we think, learn, and act. In our model, efferent copy is routed back as an input to the deep network, allowing it to learn representations that multiplex the structure and contents of a visual scene. Our work thus draws upon a long tradition from psychology and neurobiology, which argues that the neural signals responsible for controlling movement play a key role in cognition.[53] Here, we focus on representation learning, describing a model that is rooted in the proposed functional architecture and connectivity of the primate dorsal and ventral stream and that is capable of learning about the structure of visual scenes even when the objects they contain are novel, just like primates can. The theory that motivates this work has already been described in Summerfield et al.[2]

This idea has some resemblance to that proposed by emerging theories of learning in the hippocampal-entorhinal system, where the structure state spaces may be learned by taking actions and observing state transitions in an allocentric frame of reference.[54] Our model also helps to unpack the puzzling relationship between space, number, and attention in the dorsal stream structures such as the PPC. We propose that neural systems learn to allocate objects to spatial locations in a visual scene by using attention to orient across a scene in structured ways, which allows agents to multiplex information about what and where, when learning new representations. By orienting attention, we can learn explicit representations of space that are not tied to a fixed set of familiar objects and that are typically found in the PPC. Although our theory emphasizes overt attention (saccades), covert attention (in which an internal spatial focus of processing is systematically oriented without a gaze movement) is likely to play a significant role in this process. Indeed, the assumption that the PPC receives information about the covertly attended location was necessary to account for the data from our eye-tracking experiment.

We studied a very limited aspect of numerical cognition, which is the ability to enumerate a small number of novel items in a visual scene. Our findings bear only tangentially on the other ways that numbers may be used, such as, for example, in mathematical calculations involving symbolic digits. Although there is evidence for parietal involvement in arithmetic, our model does not attempt to capture this ability, which (unlike visual enumeration) is unique to humans and involves additional learning about the meaning of numerical symbols.[55] Nor is our model optimized to capture judgements about approximate number that can

be made when an array contains tens or even hundreds of items.[56,57] As a serial model, our work does not explain the constant reaction times for very small numbers (1–3 items) in the absence of distracters, which may be better captured by pattern matching mechanisms in the ventral stream.[58] We note, however, that, consistent with our model, reaction times and number of saccades have been shown to increase with number of items for the range tested in our human experiment (3–6 items),[41] and a recent systematic review and meta-analysis has established that visual attention is integral in subitizing.[59]

Many previous investigations of visual numerosity in deep neural networks have primarily been concerned with an innate or intuitive sense of visual number, which allows many animals and human infants to discriminate sets with different numerosity, even without explicit training. In neural networks, this "number sense" has been identified with the presence of "number-detector" units that respond to number while being relatively insensitive to other features, like the shape, size, and spacing of the items. Such number detectors have been found in networks trained on visual object recognition,[60,61] unsupervised objectives,[62,63] action prediction,[64] or networks that are not trained at all (randomly initialized networks).[65,66] This body of work has primarily been concerned with innate neural circuitry that is hypothesized to serve approximate numerical comparisons up to 20 or 30 items. In our work, on the other hand, we explicitly train our networks to perform exact enumeration of small numbers and we are primarily interested in the human ability to generalize systematically after learning to count—recognizing learned relations (numerosity) in novel scenes. Innate number detectors are insufficient to explain this "structure learning," which requires experience in humans and is not displayed by standard deep networks trained to classify exact numerosity.[17,18,67,68] Moreover, recent work has shown that number-detector units present before training were not critical to the formation of the number representations observed post training.[69] However, aspects of our model find support in recent related works that emphasize the role of recurrent computation in the PPC and object recognition pretraining in the emergence of numerical representations and behavior.[69,70]

The experiments and stimuli described here are deliberately made simple. Our goal is not to solve large-scale engineering challenges in computer vision but to use deep networks as a vehicle for implementing a principle from neuroscience and show how it can explain neural and behavioral phenomena observed in biological systems. Nevertheless, we believe that the principles described here could be scaled and may be useful for AI research. Indeed, glimpsing neural networks have previously been applied to tasks like visual object recognition.[23,24,71] One key outstanding question, which we leave for future work, is whether the approach described here could be used to help counting in naturalistic images (e.g., three-dimensional [3D] tabletop scenes[72]). It would also be very interesting to study zero-shot estimation of continuous quantity, such as the relative height or volume of novel shapes in a scene, and whether the model produces well-known biases, such as scalar variability (Weber-like compression). Another question is whether our approach extends to other Gestalt principles and can be deployed to explain the human ability to judge relations of proximity, similarity, enclosure, symmetry, and continuity with wholly novel objects. Finally, we focus on eye movements, but the principle described here is more general and could, in theory, extend to reaching movements, which no doubt help teach children about the structure of peri-personal space. Indeed, manual pointing in children seems to play an important role in learning[73,74] and has previously been employed in an RNN model of counting.[75]

Finally, we caveat our work with the recognition that our model does not aspire to offer a complete description of how biological agents solve the problem of visual counting. The model itself is a simplification of the mechanisms that are most likely occurring in the brain and is unable to capture them exactly. As is typically the case in a computational modeling study, there may be other modeling choices, not explored here, which offer a comparable or better account of relevant observations. However, our model is able to capture a wide range of different empirical phenomena (relating to both behavior and brain activity) observed during visual counting and describes a candidate mechanism that is inspired by an understanding of the functional neurophysiology and connectivity in the primate brain.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Jessica Thompson (jessica.thompson@psy.ox.ac.uk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

De-identified human behavior data have been deposited at Open Science Framework and are publicly available as of the date of publication at https://www.doi.org/10.17605/OSF.IO/H6EVT.

All original code has been deposited at Github and is publicly available at https://github.com/summerfieldlab/saccades as of the date of publication.

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.

AUTHOR CONTRIBUTIONS

Conceptualization, C.S., M.P., J.A.F.T., and H.S.; methodology, J.A.F.T., T.D., and H.S.; investigation, J.D.S., J.A.F.T., and T.D.; writing—original draft, C.S., J.A.F.T., and J.D.S.; writing—review and editing, J.A.F.T., C.S., T.D., J.D.S., and M.P.; funding acquisition, C.S.; resources, C.S.; supervision, C.S. and J.A.F.T.

DECLARATION OF INTERESTS

At the time of submission, C.S. was an employee of the UK Artificial Intelligence Safety Institute and H.S. was an employee of Google Deepmind.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.neuron.2024.10.008.

Received: May 9, 2024
Revised: August 6, 2024
Accepted: October 6, 2024
Published: November 1, 2024

REFERENCES

1. Wagemans, J., Elder, J.H., Kubovy, M., Palmer, S.E., Peterson, M.A., Singh, M., and von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. Psychol. Bull. *138*, 1172–1217. https://doi.org/10.1037/a0029333.
2. Summerfield, C., Luyckx, F., and Sheahan, H. (2020). Structure learning and the posterior parietal cortex. Prog. Neurobiol. *184*, 101717. https://doi.org/10.1016/j.pneurobio.2019.101717.
3. Riddoch, M.J., and Humphreys, G.W. (2003). Visual agnosia. Neurol. Clin. *21*, 501–520. https://doi.org/10.1016/S0733-8619(02)00095-6.
4. Tanaka, K. (1996). Inferotemporal cortex and object vision. Annu. Rev. Neurosci. *19*, 109–139. https://doi.org/10.1146/annurev.ne.19.030196.000545.
5. DiCarlo, J.J., Zoccolan, D., and Rust, N.C. (2012). How does the brain solve visual object recognition? Neuron *73*, 415–434. https://doi.org/10.1016/j.neuron.2012.01.010.
6. Yamins, D.L.K., and DiCarlo, J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. *19*, 356–365. https://doi.org/10.1038/nn.4244.
7. Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. Annu. Rev. Vis. Sci. *1*, 417–446. https://doi.org/10.1146/annurev-vision-082114-035447.
8. Lindsay, G.W. (2021). Convolutional neural networks as a model of the visual system: past, present, and future. J. Cogn. Neurosci. *33*, 2017–2031. https://doi.org/10.1162/jocn_a_01544.
9. Doerig, A., Sommers, R.P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G.W., Kording, K.P., Konkle, T., van Gerven, M.A.J., Kriegeskorte, N., et al. (2023). The neuroconnectionist research programme. Nat. Rev. Neurosci. *24*, 431–450. https://doi.org/10.1038/s41583-023-00705-w.
10. Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. USA *111*, 8619–8624. https://doi.org/10.1073/pnas.1403112111.
11. Cichy, R.M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. Nat. Neurosci. *17*, 455–462. https://doi.org/10.1038/nn.3635.
12. Kietzmann, T.C., Spoerer, C.J., Sörensen, L.K.A., Cichy, R.M., Hauk, O., and Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. Proc. Natl. Acad. Sci. USA *116*, 21854–21863. https://doi.org/10.1073/pnas.1905544116.
13. Güçlü, U., and van Gerven, M.A.J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J. Neurosci. *35*, 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015.
14. Bottini, R., and Doeller, C.F. (2020). Knowledge across reference frames: cognitive maps and image spaces. Trends Cogn. Sci. *24*, 606–619. https://doi.org/10.1016/j.tics.2020.05.008.
15. O'Reilly, R.C., Ranganath, C., and Russin, J.L. (2022). The structure of systematicity in the brain. Curr. Dir. Psychol. Sci. *31*, 124–130. https://doi.org/10.1177/09637214211049233.
16. Kim, J., Ricci, M., and Serre, T. (2018). Not-So-CLEVR: learning same–different relations strains feedforward neural networks. Interface Focus *8*, 20180011. https://doi.org/10.1098/rsfs.2018.0011.
17. Wu, X., Zhang, X., and Shu, X. (2019). Cognitive deficit of deep learning in numerosity. AAAI *33*, 1303–1310. https://doi.org/10.1609/aaai.v33i01.33011303.
18. Zhang, X., and Wu, X. (2020). On numerosity of deep neural networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.2011.08674.
19. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. Preprint at arXiv. https://doi.org/10.48550/arXiv.2204.06125.
20. Kinniment, M., Sato, L.J.K., Du, H., Goodrich, B., Hasin, M., Chan, L., Miles, L.H., Lin, T.R., Wijk, H., Burget, J., et al. (2024). Evaluating language-model agents on realistic autonomous tasks. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.11671.
21. Jagadeesh, A.V., and Gardner, J.L. (2022). Texture-like representation of objects in human visual cortex. Proc. Natl. Acad. Sci. USA *119*, e2115302119. https://doi.org/10.1073/pnas.2115302119.
22. Geirhos, R., Schütt, H.H., Temme, C.R.M., Bethge, M., Rauber, J., and Wichmann, F.A. (2018). Generalisation in humans and deep neural networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1808.08750.
23. Larochelle, H., and Hinton, G. (2010). Learning to combine foveal glimpses with a third-order Boltzmann machine. In Advances in Neural Information Processing Systems 23 (NIPS 2010). https://doi.org/10.1016/S0196-9781(02)00145-6.
24. Mnih, V., Heess, N., Graves, A., and Kavukcuoglu, K. (2014). Recurrent models of visual attention. Preprint at arXiv. https://doi.org/10.48550/arXiv.1406.6247.
25. Gold, J.I., and Shadlen, M.N. (2007). The neural basis of decision making. Annu. Rev. Neurosci. *30*, 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038.
26. Cloutman, L.L. (2013). Interaction between dorsal and ventral processing streams: where, when and how? Brain Lang. *127*, 251–263. https://doi.org/10.1016/j.bandl.2012.08.003.
27. Rolls, E.T., Deco, G., Huang, C.-C., and Feng, J. (2023). The human posterior parietal cortex: effective connectome, and its relation to function. Cereb. Cortex *33*, 3142–3170. https://doi.org/10.1093/cercor/bhac266.
28. van Polanen, V., and Davare, M. (2015). Interactions between dorsal and ventral streams for controlling skilled grasp. Neuropsychologia *79*, 186–191. https://doi.org/10.1016/j.neuropsychologia.2015.07.010.
29. Krauzlis, R.J., Lovejoy, L.P., and Zénon, A. (2013). Superior colliculus and visual spatial attention. Annu. Rev. Neurosci. *36*, 165–182. https://doi.org/10.1146/annurev-neuro-062012-170249.
30. Berman, R.A., and Wurtz, R.H. (2010). Functional identification of a pulvinar path from superior colliculus to cortical area MT. J. Neurosci. *30*, 6342–6354. https://doi.org/10.1523/JNEUROSCI.6176-09.2010.
31. Berman, R.A., and Wurtz, R.H. (2011). Signals conveyed in the pulvinar pathway from superior colliculus to cortical area MT. J. Neurosci. *31*, 373–384. https://doi.org/10.1523/JNEUROSCI.4738-10.2011.
32. Lyon, D.C., Nassi, J.J., and Callaway, E.M. (2010). A disynaptic relay from superior colliculus to dorsal stream visual cortex in macaque monkey. Neuron *65*, 270–279. https://doi.org/10.1016/j.neuron.2010.01.003.

33. Takayama, Y., Sugishita, M., Akiguchi, I., and Kimura, J. (1994). Isolated acalculia due to left parietal lesion. Arch. Neurol. *51*, 286–291. https://doi.org/10.1001/archneur.1994.00540150084021.

34. Ashkenazi, S., Henik, A., Ifergane, G., and Shelef, I. (2008). Basic numerical processing in left intraparietal sulcus (IPS) acalculia. Cortex *44*, 439–448. https://doi.org/10.1016/j.cortex.2007.08.008.

35. Benavides-Varela, S., Piva, D., Burgio, F., Passarini, L., Rolma, G., Meneghello, F., and Semenza, C. (2017). Re-assessing acalculia: distinguishing spatial and purely arithmetical deficits in right-hemisphere damaged patients. Cortex *88*, 151–164. https://doi.org/10.1016/j.cortex.2016.12.014.

36. Buckley, M.J., and Gaffan, D. (1998). Perirhinal cortex ablation impairs configural learning and paired–associate learning equally. Neuropsychologia *36*, 535–546. https://doi.org/10.1016/S0028-3932(97)00120-6.

37. Buckley, M.J., and Gaffan, D. (2006). Perirhinal cortical contributions to object perception. Trends Cogn. Sci. *10*, 100–107. https://doi.org/10.1016/j.tics.2006.01.008.

38. Sarnecka, B.W., and Carey, S. (2008). How counting represents number: what children must learn and when they learn it. Cognition *108*, 662–674. https://doi.org/10.1016/j.cognition.2008.05.007.

39. DeWind, N.K., Bonner, M.F., and Brannon, E.M. (2020). Similarly oriented objects appear more numerous. J. Vision *20*, 4. https://doi.org/10.1167/jov.20.4.4.

40. Friedman-Hill, S.R., Robertson, L.C., and Treisman, A. (1995). Parietal contributions to visual feature binding: evidence from a patient with bilateral lesions. Science *269*, 853–855. https://doi.org/10.1126/science.7638604.

41. Watson, D.G., Maylor, E.A., and Bruce, L.A.M. (2007). The role of eye movements in subitizing and counting. J. Exp. Psychol. Hum. Percept. Perform. *33*, 1389–1399. https://doi.org/10.1037/0096-1523.33.6.1389.

42. Nieder, A., and Dehaene, S. (2009). Representation of number in the brain. Annu. Rev. Neurosci. *32*, 185–208. https://doi.org/10.1146/annurev.neuro.051508.135550.

43. Roitman, J.D., Brannon, E.M., and Platt, M.L. (2012). Representation of numerosity in posterior parietal cortex. Front. Integr. Neurosci. *6*, 25. https://doi.org/10.3389/fnint.2012.00025.

44. Viswanathan, P., and Nieder, A. (2013). Neuronal correlates of a visual "sense of number" in primate parietal and prefrontal cortices. Proc. Natl. Acad. Sci. USA *110*, 11187–11192. https://doi.org/10.1073/pnas.1308141110.

45. Nieder, A., Diester, I., and Tudusciuc, O. (2006). Temporal and spatial enumeration processes in the primate parietal cortex. Science *313*, 1431–1435. https://doi.org/10.1126/science.1130308.

46. Karami, A., Castaldi, E., Eger, E., and Piazza, M. (2023). Neural codes for visual numerosity independent of other quantities are present both in the dorsal and in the ventral stream of the human brain. Preprint at bioRxiv. https://doi.org/10.1101/2023.12.18.571155.

47. Nieder, A., and Miller, E.K. (2004). A parieto-frontal network for visual numerical information in the monkey. Proc. Natl. Acad. Sci. USA *101*, 7457–7462. https://doi.org/10.1073/pnas.0402239101.

48. Roitman, J.D., Brannon, E.M., and Platt, M.L. (2007). Monotonic coding of numerosity in macaque lateral intraparietal area. PLoS Biol. *5*, e208. https://doi.org/10.1371/journal.pbio.0050208.

49. Barnett, B., and Fleming, S.M. (2024). Creating something out of nothing: symbolic and non-symbolic representations of numerical zero in the human brain. Preprint at bioRxiv. https://doi.org/10.1101/2024.01.30.577906.

50. Colby, C.L., and Goldberg, M.E. (1999). Space and attention in parietal cortex. Annu. Rev. Neurosci. *22*, 319–349. https://doi.org/10.1146/annurev.neuro.22.1.319.

51. Shadlen, M.N., and Newsome, W.T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. J. Neurophysiol. *86*, 1916–1936. https://doi.org/10.1152/jn.2001.86.4.1916.

52. Gnadt, J.W., and Breznen, B. (1996). Statistical analysis of the information content in the activity of cortical neurons. Vision Res. *36*, 3525–3537. https://doi.org/10.1016/0042-6989(96)00049-1.

53. Wilson, M. (2002). Six views of embodied cognition. Psychon. Bull. Rev. *9*, 625–636. https://doi.org/10.3758/BF03196322.

54. Whittington, J.C.R., Muller, T.H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T.E.J. (2019). The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalisation in the Hippocampal Formation. Preprint at bioRxiv. https://doi.org/10.1101/770495.

55. Dehaene, S. (2011). The Number Sense: How the Mind Creates Mathematics, Revised and Updated Edition Updated Edition, Updated Edition (Oxford University Press).

56. Piazza, M., Izard, V., Pinel, P., Le Bihan, D., and Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. Neuron *44*, 547–555. https://doi.org/10.1016/j.neuron.2004.10.014.

57. Cheyette, S.J., and Piantadosi, S.T. (2019). A primarily serial, foveal accumulator underlies approximate numerical estimation. Proc. Natl. Acad. Sci. USA *116*, 17729–17734. https://doi.org/10.1073/pnas.1819956116.

58. Kutter, E.F., Dehnen, G., Borger, V., Surges, R., Mormann, F., and Nieder, A. (2023). Distinct neuronal representation of small and large numbers in the human medial temporal lobe. Nat. Hum. Behav. *7*, 1998–2007. https://doi.org/10.1038/s41562-023-01709-3.

59. Chen, J., Paul, J.M., and Reeve, R. (2022). Manipulation of attention affects subitizing performance: A systematic review and meta-analysis. Neurosci. Biobehav. Rev. *139*, 104753. https://doi.org/10.1016/j.neubiorev.2022.104753.

60. Nasr, K., Viswanathan, P., and Nieder, A. (2019). Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. Sci. Adv. *5*, eaav7903. https://doi.org/10.1126/sciadv.aav7903.

61. Zhou, C., Xu, W., Liu, Y., Xue, Z., Chen, R., Zhou, K., and Liu, J. (2021). Numerosity representation in a deep convolutional neural network. J. Pac. Rim Psychol. *15*. https://doi.org/10.1177/18344909211012613.

62. Stoianov, I., and Zorzi, M. (2012). Emergence of a "visual number sense" in hierarchical generative models. Nat. Neurosci. *15*, 194–196. https://doi.org/10.1038/nn.2996.

63. Testolin, A., Dolfi, S., Rochus, M., and Zorzi, M. (2020). Visual sense of number vs. sense of magnitude in humans and machines. Sci. Rep. *10*, 10045. https://doi.org/10.1038/s41598-020-66838-5.

64. Kondapaneni, N., and Perona, P. (2024). A number sense as an emergent property of the manipulating brain. Sci. Rep. *14*, 6858. https://doi.org/10.1038/s41598-024-56828-2.

65. Park, J., and Huber, D.E. (2022). A visual sense of number emerges from divisive normalization in a simple center-surround convolutional network. eLife *11*, 1–16. https://doi.org/10.7554/eLife.80990.

66. Kim, G., Jang, J., Baek, S., Song, M., and Paik, S.B. (2021). Visual number sense in untrained deep neural networks. Sci. Adv. *7*, 1–9. https://doi.org/10.1126/sciadv.abd6127.

67. Pecyna, L., Cangelosi, A., and Nuovo, A.D. (2019). A deep neural network for finger counting and numerosity estimation. In IEEE Symposium Series on Computational Intelligence (SSCI) 2019, pp. 1422–1429. https://doi.org/10.1109/SSCI44817.2019.9002694.

68. Creatore, C., Sabathiel, S., and Solstad, T. (2021). Learning exact enumeration and approximate estimation in deep neural network models. Cognition *215*, 104815. https://doi.org/10.1016/j.cognition.2021.104815.

69. Mistry, P.K., Strock, A., Liu, R., Young, G., and Menon, V. (2023). Learning-induced reorganization of number neurons and emergence of numerical representations in a biologically inspired neural network. Nat. Commun. *14*, 3843. https://doi.org/10.1038/s41467-023-39548-5.

70. Verma, B.K., and SenGupta, R. (2023). Emergence of behavioral phenomena and adaptation effects in human numerosity decoder using recurrent neural networks. Sci. Rep. *13*, 19571. https://doi.org/10.1038/s41598-023-44535-3.

71. Adeli, H., Ahn, S., and Zelinsky, G.J. (2023). A brain-inspired object-based attention network for multiobject recognition and visual reasoning. J. Vision *23*, 16. https://doi.org/10.1167/jov.23.5.16.

72. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., and Girshick, R. (2017). CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In 30th IEEE Conference on Computer Vision and Pattern Recognition (CPVR), pp. 1988–1997. https://doi.org/10.1109/CVPR.2017.215.

73. Lucca, K., and Wilbourn, M.P. (2018). Communicating to learn: infants' pointing gestures result in optimal learning. Child Dev. *89*, 941–960. https://doi.org/10.1111/cdev.12707.

74. Coccoz, V., Lozada, M., Salsa, A., and Scheuer, N. (2019). Enactive experience promotes early number understanding: a study with 3-year-old children. J. Cogn. Psychol. *31*, 891–901. https://doi.org/10.1080/20445911.2019.1676758.

75. Fang, M., Zhou, Z., Chen, S., and McClelland, J.L. (2018). Can a Recurrent Neural Network Learn to Count Things? In Proceedings of the 40th Annual Conference of the Cognitive Science Society, pp. 360–365.

76. Grill-Spector, K., and Malach, R. (2004). The human visual cortex. Annu. Rev. Neurosci. *27*, 649–677. https://doi.org/10.1146/annurev.neuro.27.070203.144220.

77. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: an imperative style, high-performance deep learning library. Preprint at arXiv. https://doi.org/10.48550/arXiv.1912.01703.

78. Loshchilov, I., and Hutter, F. (2019). Decoupled weight decay regularization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1711.05101.

79. Thompson, J.A.F., Sheahan, H., and Summerfield, C. (2022). Learning to count visual objects by combining "what" and "where" in recurrent memory. In Proceedings of the 1st Gaze Meets ML Workshop, PMLR, pp. 199–218.

80. Brainard, D.H. (1997). The psychophysics toolbox. Spat. Vision *10*, 433–436. https://doi.org/10.1163/156856897X00357.

Please cite this article in press as: Thompson et al., Zero-shot counting with a dual-stream neural network model, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.10.008

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Human Behavioural Data	This paper	https://www.doi.org/10.17605/OSF.IO/H6EVT
Software and algorithms		
Custom code for Simulations, Experiments and Data Analysis	This paper	https://www.doi.org/10.5281/zenodo.13860802
Python 3.8.10	Python Software Foundation	https://www.python.org/; RRID: SCR_008394
MATLAB R2023a	Mathworks	https://uk.mathworks.com/products/matlab.html; RRID:SCR_001622
Psychtoolbox 3	Mario Kleiner, David Brainard, Denis Pelli, Chris Broussard, Tobias Wolf, Diederick Niehorster	http://psychtoolbox.org/; RRID: SCR_002881
Numpy 1.23.4	Community Project	https://numpy.org/; RRID: SCR_008633
Scipy 1.10.1	Community Project	https://scipy.org/; RRID: SCR_008058
Scikit-Image 0.19.2	Community Project	https://scikit-image.org/; RRID: SCR_021142
Pytorch 1.12.1	Meta	https://pytorch.org/; RRID: SCR_018536

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human participants

Twenty-six participants (age 24.96 ± 2.69; 42% female, 58% male) with normal or corrected to normal vision took part. All of them were students or researchers at the University of Oxford. They received £15 as compensation for their time (60–90 minutes). Two participants were excluded from the subsequent analyses due to data corruption. This study received ethical approval from the Central University Research Ethics Committee of the University of Oxford. All participants provided informed written consent.

METHOD DETAILS

Neural network simulations

Stimuli and task sets

We synthesised grayscale images (42 x 48 pixels) containing 1–5 target items (alphanumeric characters). For the "ignore distracters" task, an additional 0–2 distracter items were included. The primary objective was simply to report the number of target items in the image, either with ("ignore distracters") or without ("simple counting") potential distracter items. Characters lay on an invisible 6 x 6 grid, were of constant size (5 pixels tall, 4 pixels wide), and were never overlapping. N items were assigned to spatial locations in the image by randomly choosing N of the 36 possible grid locations. All target items within one image were the same character and had the same mean luminance value. Gaussian noise with standard deviation of 0.05 was added to both the background and foreground pixels.

For each task, we generated five datasets—one for training and four for testing. These datasets are summarized in Table 1. The five number classes were evenly represented in each dataset (and perfectly crossed with number of distracters in "ignore distracters"). The target shape, mean background luminance, and mean foreground luminance were sampled randomly from the set of target characters and set of mean luminances for that dataset. In the training set, target shapes were sampled from {B, C, D, E} and mean luminances from {0.1, 0.4, 0.7}. In any test set, these stimulus parameters were either the same as in the training or sampled from non-overlapping sets {F, G, H, J} and {0.3, 0.6, 0.9}. The distracter shape was always the character 'A'.

Glimpsing

Like the primate visual system, our dual-stream RNN apprehends an image via a sequence of foveated glimpses. A sequence of fixation points is generated according to a fixed, saliency-based saccadic policy. The saliency map is composed from a mixture of Gaussians with one Gaussian centred on each item. Our saccadic policy samples fixation points from the saliency map, subject to the constraint that each Gaussian is sampled at least once (i.e., each item is glimpsed at least once). To tune our saccadic policy, we used human eye tracking data from an independent study which presented similar images containing alphanumeric characters. First, we set our number of fixations to 12 based on the observation that, during a 3 second viewing period, human participants made 12 saccades on average. Second, we matched the standard deviation (in both x and y directions) of the fixation coordinates in the

Please cite this article in press as: Thompson et al., Zero-shot counting with a dual-stream neural network model, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.10.008

reference frame of the nearest item. In the human data, this standard deviation was 5% of the total image width/height. We adjusted the dispersion of the Gaussians in the saliency map to replicate this property in the simulated fixations (Figure S8).

The mapping from the retina to the cortex in humans is well described as a log-polar transformation in which the horizontal and vertical axes in the retina are transformed into polar axes in the cortex: angle and eccentricity (distance from fovea—log scaled).[76] To prepare our glimpse contents, we simulated the retinal-to-cortical transformation as a log-polar transform centred on the fixation coordinate using the *warp_polar* function from the Scikit-Image python package (version 0.19.2). The radius of the circle that bounded the transformed area was $r = \sqrt{width^2+height^2}$ such that the entire image would be included in a glimpse directed at the centre of the image. Points outside the boundaries of the image were filled according to the 'edge' interpolation mode which pads missing values with neighbouring values in the image. The output shape of the log-polar transform was set to be the same size as the original image. The log scaling of the eccentricity coordinates results in foveated glimpse contents in which the foveal region is magnified relative to the periphery.

Model architecture

Our dual-stream RNN, inspired by the parallel pathways of the primate visual system, receives both the glimpse positions (the fixation coordinates) and the glimpse contents (the log-polar transformed image) as separate inputs streams. Separate feed-forward layers produce equal-sized embeddings (512 units) of both glimpse positions and contents. These two embeddings are concatenated before being passed through another layer to produce a joint embedding (1024 units). For the "ignore distracters" task, the contents embedding layer is preceded by a convolutional module consisting of three convolutional layers and two feedforward layers. The activations of the penultimate ventral layer are passed on to the contents embedding layer. The joint embedding layer is followed by a recurrent submodule. The recurrent submodule consists of three transformations (input to hidden, hidden to hidden, and hidden to out) which preserve the size of the representation (1024). The output of the recurrent submodule is passed to a feed forward layer with 36 units. We call this layer the 'map layer' because we train it to reflect the spatial arrangement of target items in the image. A final linear readout layer (5 units) predicts the number of target items in the image. For the "simple counting" task, this results in 5,268,696 total trainable parameters. All activations functions are leaky rectified linear (slope=0.1) except for on the map layer where a sigmoid activation is used to get values between 0 and 1. The model architecture is depicted in Figure 3B, and detailed model parameters are listed in Tables S1 and S2. This design embodies the hypothesis that efferent copies of signals pertaining to a viewer's orientation to a scene, e.g., eye movements, rather than merely outputs of a visual system are also inputs, that in turn support learning useful representations of space and number.

Our convolutional neural network baseline model consists of three convolutional layers and two fully connected layers before the number readout. All convolutional layers consist of 56 feature maps, use a stride of 1, and use no padding. The first convolutional layer uses a 3x3 kernel and subsequent convolutional layers use a 2x2 kernel. The first fully connected layer has 256 units, and the last layers have 36 and 5 units respectively. This results in 5,454,461 total trainable parameters. Detailed network parameters are listed in Table S3.

Model training

All networks were built and trained using the PyTorch python library version 1.12.1[77] on a single NVIDIA Titan X Pascal GPU. Additional python packages NumPy 1.23.4, Matplotlib 3.5.3, Pandas 1.5.3, Seaborn 0.13.2, and SciPy 1.10.1 were used for data analysis and visualisation.

For the "ignore distracters" task, the total set of trainable model parameters of the dual-stream RNN θ can be divided into those that make up the convolutional (or ventral) module $\mathcal{V}(c_g)$ where c_g are the glimpse contents of a particular glimpse g and those that make up the recurrent (or parietal) module $\mathcal{P}(\mathcal{V}_{-1}(\boldsymbol{c}), \boldsymbol{p})$, where $\mathcal{V}_{-1}$ indicates the output of the penultimate layer of the convolutional module and $\boldsymbol{c}$ and $\boldsymbol{p}$ are the sequence of 12 glimpse contents and positions respectively.

$$\theta = \{\theta_{ventral}, \theta_{parietal}\}$$

These parameters are optimised with respect to three different objective functions: a shape recognition loss $\mathcal{L}_{shape}(\theta_{ventral})$, a spatial map loss $\mathcal{L}_{map}(\theta_{parietal})$, and a number classification loss $\mathcal{L}_{number}(\theta_{parietal})$. During a pretraining phase, the convolutional module is trained to predict the proximity of a glimpse to any nearby target or distracter items.

$$\mathcal{L}_{shape}(\theta_{\text{ventral}}) = \text{MSE}(s_g, \mathcal{V}(\mathbf{c_g}))$$

The target vector $\boldsymbol{s_g}$ for this shape recognition task is constructed as follows. For a particular glimpse g, we calculate the Euclidean distances d_i from the fixation point to every item i in the image within 3σ of the fixation point (where the dispersion of the isotropic Gaussians that make up the saliency map is σ^2). For each item within range, we calculate the proximity as $1 - d_i/3\sigma$. The final target vector $\boldsymbol{s}$ consists of two values which correspond to the sum of the proximities for distracters and targets respectively. For example, a glimpse directed exactly at the centre of a target item with no other items in the vicinity would produce a proximity vector $\boldsymbol{s_g}$ of [0, 1] indicating that this is a very 'targety' glimpse. A similar vector would also obtain for a glimpse directed in between two neighbouring target items with no other items in the vicinity. If instead the glimpse was directed between a target and a distracter item, the proximity vector would be approximately [0.5, 0.5]. The convolutional module is pretrained on an independent dataset containing all letters and all luminances. The parameters of the convolutional module are held fixed during subsequent training of the recurrent module.

Please cite this article in press as: Thompson et al., Zero-shot counting with a dual-stream neural network model, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.10.008

Recall that the penultimate layer of the dual-stream RNN has 36 units corresponding to the 36 image "slots"—the spatial locations spanned by the 6x6 grid where items may appear in the image. This map layer is supervised with a binary cross entropy loss to produce a binary map of where the target items appear in the image:

$$\mathcal{L}_{map}(\theta_{parietal}) = \mathrm{BCE}(\mathbf{m}, \mathcal{P}_{-1}(\mathcal{V}_{-1}(\boldsymbol{c}), \boldsymbol{p}))$$

where the map target vector $\boldsymbol{m}$ contains a 1 for every slot that contains a target item and a 0 otherwise and $\mathcal{P}_{-1}(\mathcal{V}_{-1}(\boldsymbol{c}), \boldsymbol{p})$ indicates the output of the penultimate layer of the recurrent module (the map layer).

From this map representation, a final readout layer produces the numerosity prediction, on which a standard cross entropy classification loss is computed:

$$\mathcal{L}_{\mathrm{number}}(\theta_{\mathrm{parietal}}) = \mathrm{CE}(\mathbf{n}, \mathcal{P}(\mathcal{V}_{-1}(\boldsymbol{c}), \boldsymbol{p}))$$

where $\mathbf{n}$ is the number of target items in the image. The optimised objective function is simply the sum of the number loss and the auxiliary map loss.

$$\mathcal{L}(\theta_{parietal}) = \mathcal{L}_{number}(\theta_{parietal}) + \mathcal{L}_{map}(\theta_{parietal})$$

All models were optimised with the AdamW optimiser (Adaptive Moment Estimation with Decoupled Weight Decay Regularization)[78] with weight decay of 1e-5. For glimpsing models, the order of glimpses was randomised anew at the beginning of each epoch to effectively augment the dataset. For recurrent models, we clipped the gradient norm at 2 to stabilise learning. All models were trained with a batch size of 512 for 300 epochs with a starting learning rate of 0.001 and a scheduler that decayed the learning rate by a factor of 0.7 every 15 epochs. All models were trained 20 times from different random initializations. Reported results are averaged over these repetitions.

Control models and ablations

Ablate contents/position. When ablating one input stream or the other, the input is simply omitted both at training and test time. In this setting, what is labelled the "joint embedding" in Figure 3B is a function of either the glimpse contents *or* the glimpse positions, not both. These are thus one-stream models.

Dual-stream no map. To interrogate the role of the spatial map representation in the penultimate layer of the dual-stream RNN, in this condition, we omit the map term $\mathcal{L}_{map}(\theta_{parietal})$, updating the weights only with respect to the number objective $\mathcal{L}_{number}(\theta_{parietal})$. Whenever the map loss is not optimised, the sigmoid nonlinearity at the map layer is replaced with a leaky rectified linear (slope=0.1) to avoid vanishing gradients.

CNN+map. In this version of the CNN baseline, we add the map loss term to the training objective. As in the dual-stream RNN, we calculate the map loss on the penultimate layer of the CNN (the second fully connected layer).

Whole image RNN. This is a one-stream, non-glimpsing control model whose architecture is identical to the Ablate position model. Instead of receiving a sequence of foveated glimpse contents, it receives the whole image repeatedly. This tests the role of recurrence alone.

No ventral pretraining. Normally the parameters of the convolutional module are pretrained on the shape recognition objective $\mathcal{L}_{shape}$ and then held fixed during the training of the parietal parameters with respect to the number of map objectives. In this control model, we do not pretrain the convolutional module and we do not update with respect to the shape objective at all. Instead, all parameters θ are updated with respect to the number and map objectives during the main training phase.

$$\mathcal{L}(\theta) = \mathcal{L}_{number}(\theta) + \mathcal{L}_{map}(\theta)$$

Streams misaligned. Here we shuffle the sequence order of the glimpse positions such that they are temporally misaligned with the glimpse contents. The positions are shuffled anew on each image so that the correspondence between particular glimpse positions and contents cannot be easily inferred. By disrupting the temporal coincidence of the two input streams, we test how much of the performance of the dual-stream RNN can be explained by merely the sum of the task-relevant information in the two streams over all glimpses rather than on the integration of each glimpse pair.

Symbolic counter

The architecture of the dual stream RNN is inspired by the premise that numerosity judgements require the integration of "what" and "where". Our theory predicts that this will be more likely to be the case for some glimpsed images than others and will depend 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. We therefore wanted to assess how performance depends on the need for integration. To this end, we developed a manual algorithm that uses a series of rules to derive numerosity from a symbolic representation of the glimpses in a simplified version of the simple counting task.

This symbolic counter is equipped with knowledge about the data generating process—it knows that glimpses are generated by items in the array (in this case, according to a truncated isotropic Gaussian), it knows items lie on a grid, and it knows the range of possible numerosities. The counter determines the numerosity by counting how many of the candidate item locations (slots) contain items. It first tries to infer this from the gaze locations alone. For each glimpse, it will enumerate the candidate slots that could have reasonably generated the glimpse. If there is only one candidate slot, this is an unambiguous glimpse. After going through all glimpses, each slot in the image will be assessed to either not contain an item, contain an item, or maybe contain an item. If at this stage, if

Please cite this article in press as: Thompson et al., Zero-shot counting with a dual-stream neural network model, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.10.008

all glimpses have been unambiguously assigned to the item slots that generated them, then the counter stops here, and the glimpsed image would receive an integration score of 1. Otherwise, the counter picks an ambiguous glimpse and inspects the glimpse contents to disambiguate. The integration score reflects how many ambiguous glimpses had to be disambiguated via the glimpse contents to infer which slots were filled. Note that the number of glimpses is constant in these simulations. 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 counter cannot handle and are omitted from the presented analysis. See Thompson et al.[79] for more description.

Human experiment

Eye tracking environment and setup

Each participant completed the study in one session that lasted between one and one-and-a-half hours. The experimenter remained in the room and recalibrated the system once after the practice trials, and then again after every other block of trials, with recalibrations taking place approximately every 10 minutes. Participants were seated in a dark room approximately 60cm away from a computer monitor (60 Hz refresh rate, 1280*1024 resolution, 17′ LCD). Participants rested their head on an adjustable chin and head rest. Their eye gaze position was monitored using SR Research EyeLink 1000 and recorded at 1000 Hz. Fixation events were detected automatically by the SR Research Software.

Stimuli

Stimuli were synthetic grey-scale images containing between three and six *targets* (either the characters C, E, F, J, K, S, U, or Z), and – in ''ignore distracters'' trials – between one and three *distracters* (A's). Background and foreground (letter) luminances were always 0.3, 0.6, or 0.9, chosen randomly apart from the constraint that the difference in luminance was fixed at 0.3. Otherwise, image generation parameters were the same as for the neural network simulations as described above under *Stimuli and task sets*. We generated independent stimuli sets for each participant and each condition. These stimuli sets consisted of 72 images in which all numerosities were represented equally (and crossed with number of distracters, where applicable). We also generated separate stimuli sets for the training blocks – one per condition per participant.

Task procedure

Visual stimuli were presented with Psychtoolbox-3[80] for MATLAB. Each stimuli set was split over two experimental blocks of 36 trials. Each experimental session consisted of four training blocks of eight trials, followed by eight experimental blocks of 36 trials for a total of 288 experimental trials per participant. The core task for each trial was to count either all letters (''simple counting'') or all letters but A's (''ignore distracters'') present in an image. Each trial was subject to one of two viewing conditions. Participants were either allowed to move their eyes freely (free gaze), or they were asked to fixate on a red cross in the middle of the screen (fixed gaze). Fixed gaze trials were only accepted if the participant maintained central fixation (within a radius of 100 pixels of the centre of the screen) for the whole duration of the stimulus presentation. Rejected trials were appended at the end of the respective block, to ensure the complete datasets for each participant. The two tasks and viewing conditions were combined according to a 2x2 design, so there were four conditions overall.

Participants received instructions before the start of each block, indicating whether they can move their eyes and whether to count all letters or ignore A's. They confirmed they had understood by pressing any key, and then the block would begin. The four practice blocks were assigned to the four conditions in the following order: (1) *simple counting, free gaze*, (2) *ignore distracters, free gaze*, (3) *simple counting, fixed gaze*, (4) *ignore distracters, fixed gaze*. For the experimental blocks, condition order was randomly assigned, subject to the constraint that each condition appeared once in the first group of four blocks, and once in the second group. Thus, every participant completed one practice- and two experimental blocks for each condition.

Each trial followed the same structure: (1) A circle converged toward a fixation point in the screen's centre within two seconds. (2) Once the point disappeared, the stimulus was presented in a 728x728 square in the centre of the screen. It remained there until either two seconds elapsed, or the participant pressed spacebar to confirm that they were ready to respond. (3) Following presentation of a *mask image* to suppress iconic memory for 0.1 seconds and a gap of 0.5 seconds, (4) the *response screen* was displayed, which asked participants to press keyboard keys 3, 4, 5, or 6 to indicate the number of target items. Their response was displayed on the screen and they were asked to confirm their response by pressing the button again. Once the participant confirmed, they either received feedback (*correct* or *incorrect*) or, if they had moved their gaze too far away from the centre of the screen in a *fixed gaze* trial, a message reporting trial rejection for 1.5 seconds. After a brief interval of 0.5 seconds, the next trial followed.

QUANTIFICATION AND STATISTICAL ANALYSIS

Neural network simulations

Pattern of errors during learning

To see how the performance of the dual-stream RNN developed through training, we inspected confusion matrices saved at several performance-based checkpoints during training: 25% correct, 50% correct, 75% correct, and the final model at convergence. The confusion matrices in Figure 5A reflect the pattern of errors made by the dual-stream RNN on the validation set of the *ignore distracters* task at the end of the epoch in which its training performance crossed the corresponding threshold, averaged over 20 repetitions with different random seed. These epoch checkpoints were insufficient to understand how the pattern of errors develops in the CNN because the CNN learns very quickly—it can easily pass through the 25% correct and 50% correct threshold within the

Please cite this article in press as: Thompson et al., Zero-shot counting with a dual-stream neural network model, Neuron (2024), https://doi.org/10.1016/j.neuron.2024.10.008

same epoch. We therefore recorded the accuracy per number class encountered on each mini-batch during training. This provided a fine-grained view of how the per-class accuracy developed during training and allowed comparison with the dual-stream RNN.

Coherence effect

To test whether our model displays the coherence effect, in which humans tend to find arrays of more similar items to be more numerous,[39] we needed to evaluate our model on different levels of item coherence. Without any large changes to the parameters of our image generation process, we obtained four different levels of item coherence by varying the number of unique item shapes within an image. At the most coherent (or least distinctive) level, all items within an image are exactly the same letter. At the least coherent (or most distinctive) level, items are as distinctive as possible, given the parameters of the dataset. Arrays of four or five items will include all four unique letters (BCDE). For arrays of less than four items, each item will be a different letter. At the two intermediate levels of coherence, the maximum number of unique items within an image is two and three respectively. The dual-stream RNN was trained on an equal mixture of all four coherence levels so that all tests would be in-distribution.

Neural coding analyses

We analysed the responses of the 1024 units in the recurrent layer of the dual-stream RNN to the 5000 images in the OOD both test set (*simple counting*) and on each of the 12 glimpses per image. All analyses were performed in MATLAB (2023a, The MathWorks Inc., Natick, MA) using custom code.

Number Coding. For each unit and on each glimpse, we calculated the mean response to each number class to see how the single-unit response to numerosity evolved over glimpses. We calculated the preferred numerosity for each unit as the number class that elicited the largest mean response over all glimpses. Units were then grouped according to their preferred numerosity and number tuning curves were calculated as the mean response over glimpses within units that preferred the same numerosity. Gaussian curves were fit to these tuning profiles using MATLAB's *fit* function in the space of linear numerosity and log numerosity. To inspect the geometry of the population response, we first calculated the pairwise Euclidean distances between the mean response to each number class (averaged over glimpses) and then applied classical multidimensional scaling (MDS) on the resulting dissimilarity matrices. This produced the 2D representation displayed in Figure 6H. To test the dimensionality of the population activity, we split the trials into two halves which were dimensionality reduced separately according to the same procedure described above except that the pairwise distances were calculated on the responses to each image rather than the mean responses to each number class. Then a multiple linear regression was trained to predict the first half of the dataset from its dimensionality reduced version. This trained model was then tested on the second half of the dataset. The dimensionality of the reduction that produces the highest variance explained provides an estimate of the dimensionality of the population activity. We repeated this process on 1000 different random splits of the dataset.

Spatial Coding. To calculate spatial response fields for each unit, we defined a tiling of the whole image minus a 5% border on all sides resulting in a 20x20 grid of 400 tiles. For each tile, we identified all glimpses whose position was within the bounds of the tile. Then, the spatial response field (RF) was calculated for each unit as the mean response to glimpses directed at each tile. We determined the 2D Gaussian of best fit for each RF by performing a grid search over the 400 candidate means and 40 standard deviations sampled along an exponential curve from 1% to 50% of the image width. We calculated the eccentricity of each RF as the Euclidean distance from the RF's mean to the centre of the image.

Human experiment

Behavioural analysis

Behavioural analyses were performed in Python version 3.10.8 (Python Software Foundation) using the Statsmodels python package version 0.14.1. For each participant, the accuracy per condition was calculated as the number of correct trials out of the total number of trials in that condition. Performance was compared across conditions using a 2x2 ANOVA (Type III) and subsequent Tukey HSD Test (n=24). For five of the participants, 1–8 (out of 288) trials were missing due to a recording error. These trials did not belong to a particular condition, and thus none of the participants' data were excluded from the analysis.