

A Number Sense as an Emergent Property of the Manipulating Brain

N. Kondapaneni and P. Perona
California Institute of Technology

December 9, 2020

Abstract

The ability to understand and manipulate numbers and quantities emerges during childhood, but the mechanism through which this ability is developed is still poorly understood. In particular, it is not known whether acquiring such a *number sense* is possible without supervision from a teacher.

To explore this question, we propose a model in which spontaneous and undirected manipulation of small objects trains perception to predict the resulting scene changes. We find that, from this task, an image representation emerges that exhibits regularities that foreshadow numbers and quantity. These include distinct categories for zero and the first few natural numbers, a notion of order, and a signal that correlates with numerical quantity. As a result, our model acquires the ability to estimate the number of objects in the scene, as well as *subitization*, i.e. the ability to recognize at a glance the exact number of objects in small scenes. We conclude that important aspects of a facility with numbers and quantities may be learned without explicit teacher supervision.

Introduction

Mathematics, one of the most distinctive expressions of human intelligence, is founded on the ability to reason about abstract entities. We are interested in the question of how humans develop an intuitive facility with numbers and quantities, and how they come to recognize numbers as an abstract property of sets of objects. There is wide agreement that innate mechanisms play a strong role in developing a *number sense* [1, 2], that naming numbers is not necessary for the perception of quantities [3, 4], and brain areas specifically involved in processing numbers have been identified [5]. As to the role of learning, we do not yet know whether developing a number sense requires a teacher’s supervision or whether unsupervised learning is sufficient.

The role of supervised learning in developing abilities that relate to the natural numbers and estimation has been recently explored using computational models. Fang et al. [6] trained a recurrent neural network to count sequentially and Sabathiel et al. [7] showed that a neural network can be trained to anticipate the actions of a teacher on three counting-related tasks – they find that specific patterns of activity in the network’s units correlate with quantities. The ability to perceive *numerosity*, i.e. a rough estimate of the number of objects in a set, was explored by Stoianov and Zorzi [8], who trained a deep network encoder to efficiently reconstruct patterns composed of dots, and found that the network developed units or ‘neurons’ that were coarsely tuned to quantity, and by Nasr et al. [9], who found the same effect in a deep neural network that was trained on visual object classification, an unrelated task. In these models the emergent quantity-sensitive units are found to be useful input to a supervised classifier that is trained to estimate numerosity. Whether supervision, which is crucial in previous works, is intrinsically necessary to learn a number sense remains an open question. We ask whether the natural numbers, as an ordered set of abstract concepts, as well as the effortless and spontaneous perception of quantity, may be learned without explicit supervision.

We explore the hypothesis that a facility with numbers and quantities may arise through unsupervised learning, and focus on the interplay of action and perception as a possible avenue for this to happen. More specifically, we explore whether perception, as it is naturally trained during object manipulation, may develop representations that support a number sense. In order to test this hypothesis we propose a model where perception learns how specific actions modify the world. The model shows that perception develops a representation of the scene which, as an emergent property, can enable the ability to manipulate numbers and estimate quantities at a glance [10, 11]. Thus, we find that a teacher is not needed.

In order to ground intuition, consider a child who has learned to pick up objects, one at a time, and let them go at will. Imagine the child sitting comfortably and playing with small toys (acorns, Legos, sea shells) which may be dropped into a bowl. We will assume that the child has already learned to perform, and tell apart, three distinct operations. The *put* (P) operation consists of picking up an object from the surrounding space and dropping it into the bowl. The *take* (T) operation consists in doing the opposite: picking up an object from the bowl and discarding it. The *shake* (S) operation consists of agitating the bowl so that the objects inside change their position randomly without falling out. Objects in the bowl may be randomly moved during put and take as well. We hypothesize that the visual system of the child is progressively becoming sensitive to the changes that are caused by manipulation [12]. As a result, the visual system is progressively trained through spontaneous play to predict (or, more precisely, post-dict) which operation took place that changed the appearance of the bowl: was it a put, a take or a shake?

A number of perceptual maps are known to arise during development, each one of which contains mechanisms that are tuned to specific scene properties, as simple as orientation [14] and boundaries [15], and as complex as faces [16] and objects [17, 18]. We propose that, while the child is playing, the visual system is being trained to use one or more such maps to build a representation that facilitates the comparison of the pair of images that are seen

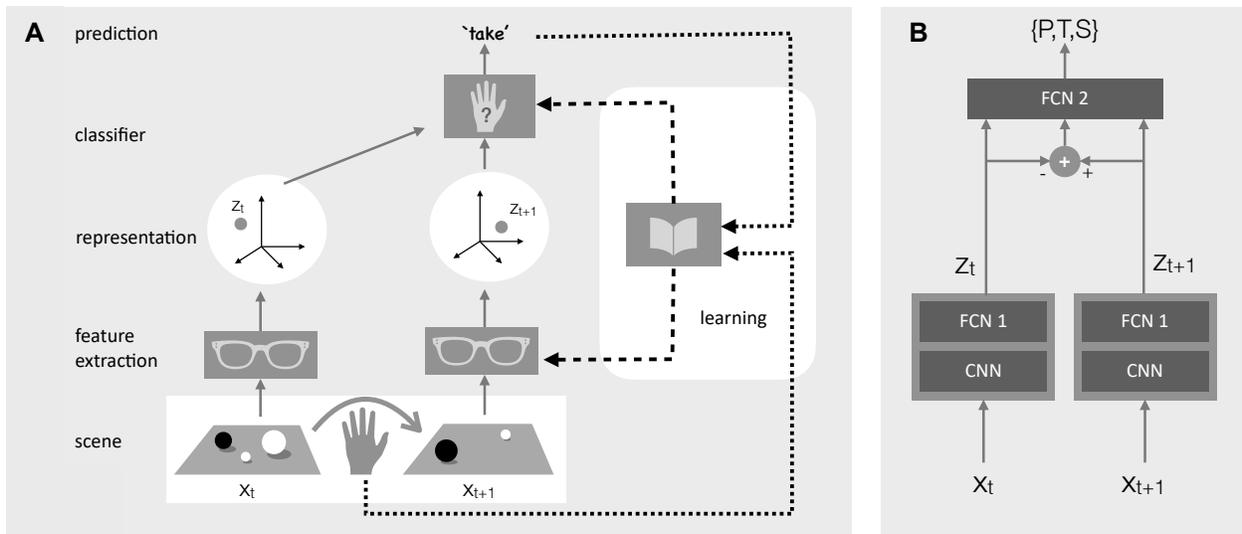


Figure 1: Schematics of our model. (A) (Bottom-to-top) The scene changes as a result of manipulation. The images x_t and x_{t+1} of the scene before and after manipulation are mapped by perception into representations z_t and z_{t+1} . These are compared by a classifier to predict which action took place. Learning monitors the error between predicted action and the efferent copy of the actual action, and updates simultaneously the weights of both perception and the classifier to increase prediction accuracy. (B) (Bottom-to-top) Our model of perception is a hybrid neural network composed of the concatenation of a convolutional neural network (CNN) with a fully-connected network (FCN 1). The classifier is implemented by a fully connected network (FCN 2) which compares the two representations z_t and z_{t+1} . The difference $z_{t+1} - z_t$ is an additional input. The two perception networks are actually the same network operating on distinct images and therefore their parameters are identical and learned simultaneously in a *Siamese network* configuration [13]. Details of the models are given in **Fig. S1**.

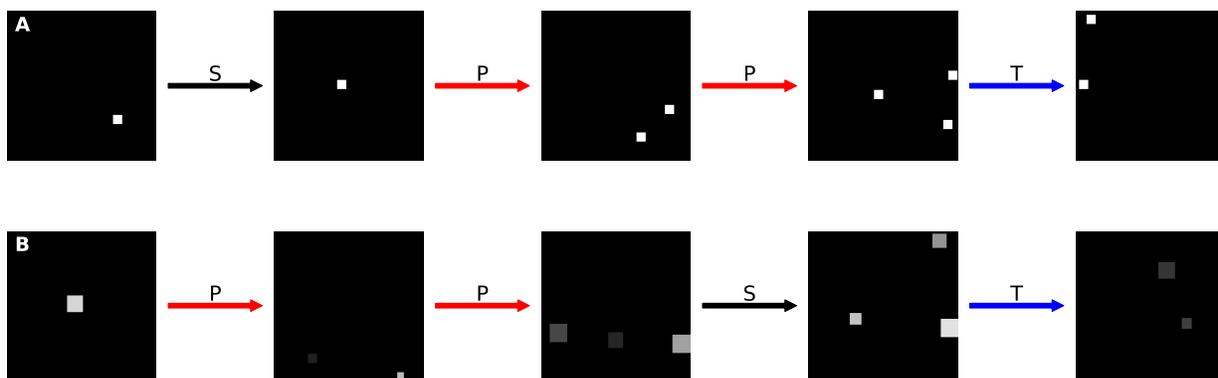


Figure 2: Training image sequence samples. We trained our model using sequences of images that were generated by randomly concatenating take (T), put (P) and shake (S) manipulations, while limiting the number of objects to the $\{0 \dots 3\}$ set (see Methods - Training Sets). We experimented with two different environment/scene statistics: (A) Identical objects (15x15 pixel squares) with random position. (B) Objects (squares) of variable position, size and contrast. The overall image intensity is a poor predictor of cardinality in this dataset (statistics in **Fig. S2**). Note that some of the objects in (B) have very low contrast and may not be visible on all displays.

before and after a manipulation. Simultaneously, a classifier network is trained to predict the action (P,T,S) from the representation of the pair of images (see **Fig. 1**). Using a simple model of this putative mechanism, we find that the image representation that is being learned for classifying actions, simultaneously learns to represent and perceive the first few natural numbers, to place them in the correct order, from zero to one and beyond, as well as estimate the number of objects in the scene.

Results

We postulate that the efferent signals from the motor system are available to the visual system and are used as a supervisory signal (**Fig. 1A**). Such signals provide information regarding the three actions of put, take and shake and, accordingly, perception may be trained to predict these three actions. Note that *no explicit signal indicating the number of objects in the scene is available to the visual system* at any time.

We use a standard deep learning model of perception [19, 20, 21]: a *feature extraction* stage is followed by a *classifier* (**Fig. 1B**). The feature extraction stage maps the image x to an internal representation z , often called an *embedding*. It is implemented by a deep network [20] composed of convolutional layers (CNN) followed by fully connected layers (FCN 1). The classifier, implemented with a simple fully connected network (FCN 2), compares the representations z_t and z_{t+1} of the *before* and *after* images to predict which action took place. Feature extraction and classification are trained jointly by minimizing the prediction error. We find that the embedding dimension makes little difference to the performance of the network (**Fig. S3**). Thus, for ease of visualization, we settled on two dimensions.

We carried out train-test experiments using sequences of synthetic images containing a small number of randomly arranged objects (**Fig. 2**). When training we limited the top number of objects to three (an arbitrary choice), and each pair of subsequent images was consistent with one of the manipulations (put, take, shake). We ran our experiments twice with different object statistics. In the first dataset the objects were identical squares, in the second they had variable size and contrast. In the following we refer to the model trained on the first dataset as *Model 1* and the model trained on the second dataset as *Model 2*.

We found that models learn to predict the three actions on a test set of novel image sequences (**Fig. 3**) with an error below 1% on scenes up to three objects (the highest number during training). Performance degrades progressively for higher numbers beyond the training range. Model 2’s error rate is higher, consistently with the task being harder. Thus, we find limited generalization of the supervised task to previously unseen numbers of objects.

When we examined the structure of the embedding we were intrigued to find a number of unexpected regularities (**Fig. 4**). First, the images’ representations do not spread across the embedding, filling the available dimensions, as is usually the case. Rather, they are arranged along a one-dimensional structure. This trait is very robust to extrapolation: after training (with up to three objects), we computed the embedding of novel images that contained up to thirty objects and found that the line-like structure persisted (**Fig. 4A**). This *embedding line* is also robust with respect to the dimensions of the embedding – we tested from two to

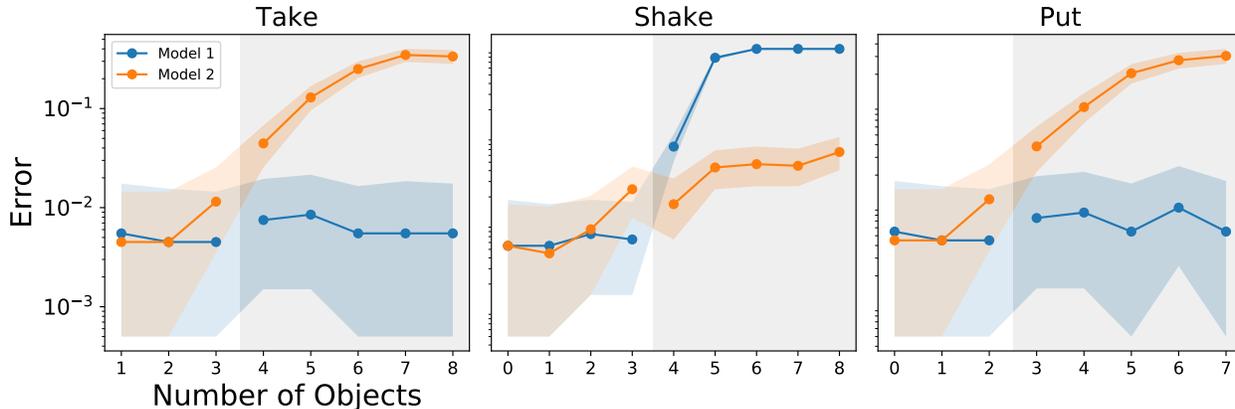


Figure 3: Action classification performance. The network accurately classifies actions up to the training limit of three objects, regardless of the statistics of the data (the x axis indicates the number of objects in the scene before the action takes place). Error increases when the number of objects in the test images exceeds the number of objects in the training set. 95% Bayesian confidence intervals are shown by the shaded areas ($272 \leq N \leq 360$). The gray region highlights test cases where the number of objects exceeds the number in the training set.

256 and observed it each time.

Second, images are arranged almost monotonically along the embedding line according to the number of objects that are present (**Fig. 4A**). Thus, the representation that is developed by the model contains an order. We were curious as to whether the *embedding coordinate*, i.e. the position of an image along the embedding line, may be used to estimate the number of objects in the image. Any one of the features that make up the coordinates of the embedding provides a handy measure for this position, measured as the distance from the beginning of the line – the value of these coordinates may be thought of as the firing rate of specific neurons [22]. We tested this hypothesis both in a relative and in an absolute quantity estimation task. First, we used the embedding coordinate to compare the number of objects in two different images and assess which is larger, and found very good accuracy (**Fig. 5**). Second, assuming that the system may self-calibrate, e.g. by using the ‘put’ action to estimate a unit of increment, then an absolute measure of quantity may be computed from the embedding coordinate. We tested this idea by computing such a *perceived number* against the actual count of objects in images (**Fig. 6**). The estimates turn out to be quite accurate, with a slight underestimate that increases as the numbers become larger. Both relative and absolute estimates of quantity were accurate as far as thirty objects (we did not test beyond this number), which far exceeds the training limit of three.

Third, image embeddings separate out into distinct ‘islands’ at one end of the embedding line (**Fig. 4A inset**). The brain is known to spontaneously cluster perceptual information [23], and therefore we tested empirically whether this form of unsupervised learning may be sufficient to discover distinct categories of images/scenes from their embedding. We found that unsupervised learning successfully discovers the clusters with very few outliers in

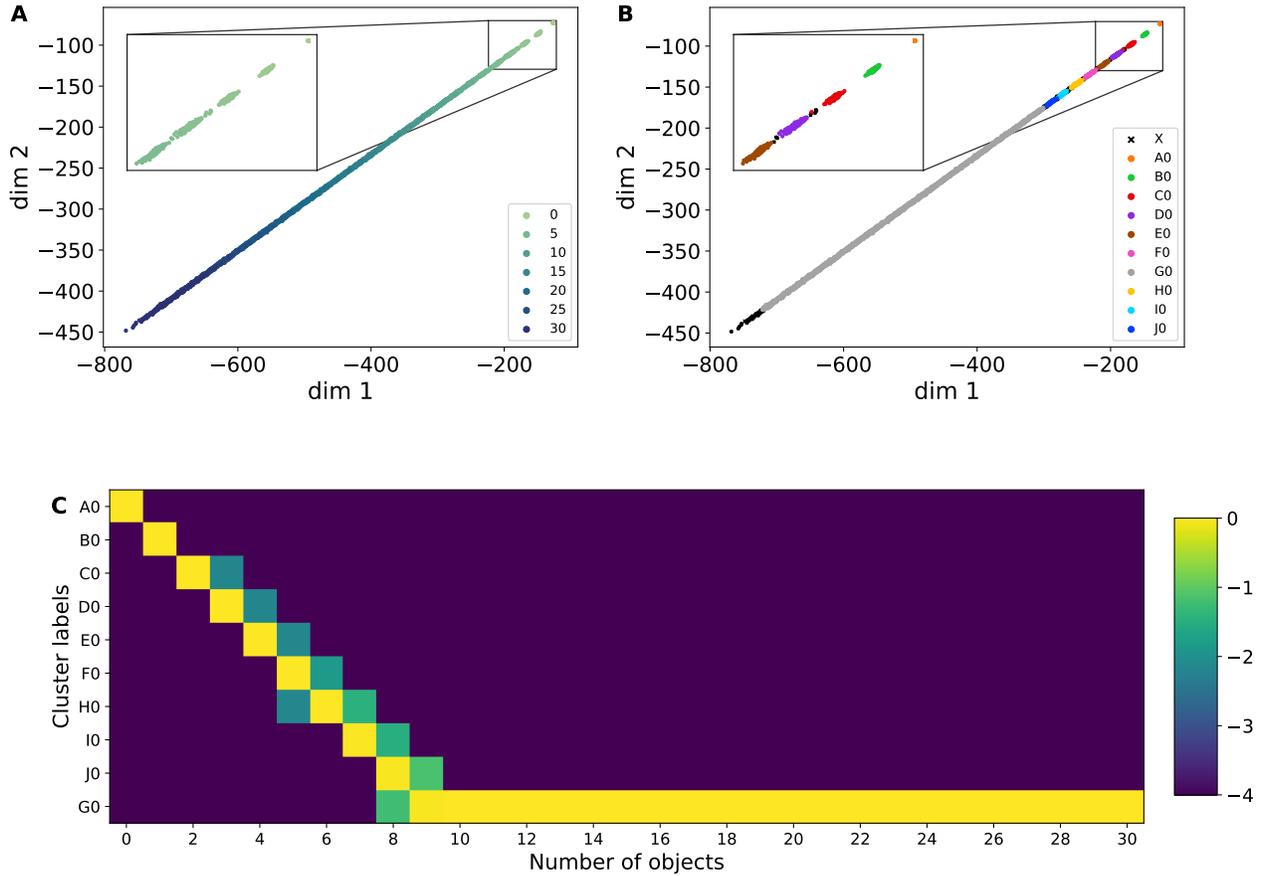


Figure 4: Visualizing the embedding space for Model 2. To explore the structure of the embedding space, we generated a dataset with $\{0 \dots 30\}$ objects, extending the number of objects far beyond the limit of 3 objects in the training task. Each image in the dataset was passed through Model 2 and the output (the internal representation/embedding) of the image is shown. **(A)** Each dot indicates an image embedding and the embeddings happen to be arranged along a line. The number of objects in each image is color coded. The smooth gradation of the color suggests that the embeddings are arranged monotonically with respect to the number of objects in the corresponding image. The inset shows that the embeddings of the images that contain only a few objects are arranged along the line into ‘islands’. **(B)** We apply an unsupervised clustering algorithm to the embeddings. Each cluster that is discovered is denoted by a specific color. The cluster X, denoted by black crosses, indicates points that the clustering algorithm excluded as outliers. **(C)** The confusion matrix shows that the clusters that are found by the clustering algorithm correspond to numbers. Images containing 0 - 8 objects are neatly separated into individual clusters, after that images are collected into a large group that is not in one-to-one correspondence with the number of objects in the image. Note that the color scale is logarithmic (base 10).

both Model 1 and the more challenging Model 2 (**Fig. 4B**).

Fourth, the first few clusters discovered by unsupervised learning along the embedding line are in almost perfect one-to-one correspondence with groups of images that share the same number of objects (**Figs. 4C**). Once such distinct *number categories* are discovered, they may be used to classify images. This is because the model maps the images to the embedding, and the unsupervised clustering algorithm can classify points in the embedding into number categories. Thus, our model learns the ability to carry out instant association of images with a small set of objects with the corresponding number category.

A fifth property of the embedding is that there is a limit to how many distinct number categories are learned. Beyond a certain number of objects, large clusters, which are no longer number-specific, form (**Fig. 4**). I.e. our model learns distinct categories for the numbers between zero and eight, and additional larger categories for, say, “more than a few” and for “many”.

There is, of course, nothing magical in the fact that during training we limited the number of objects to three – the regularities we observed are robust with respect to the choice of the number of objects that are used in training the action classifier (**Fig. S5, S6**).

Discussion

Our model and experiments demonstrate that a number sense may be learned without explicit supervision by an agent that freely engages in object manipulation. The two mechanisms of the model, deep learning and unsupervised clustering, are computational abstractions of mechanisms that have been documented in the brain. The model observes the effect of manipulation on sets of objects and learns to predict actions. The image representation that is developed in the process presents regularities which confer to the model a number of emergent properties.

First, the model discovers the structure underlying the integers. The first few numbers, from zero to eight, say, emerge as categories from spontaneous clustering of the embeddings of the corresponding images, and these number categories are naturally ordered by position on the embedding line. Remarkably, more number categories emerge than the number of objects that was present in the training set. The ability to think about numbers may be thought as a necessary, although not sufficient, step towards counting, addition and subtraction [28, 29]. The dissociation between familiarity with the first few numbers and the ability to count has been observed in hunter-gatherer societies [4] suggesting that these are distinct steps in cognition.

Second, the emergence of number categories in the embedding enables instant classification of the number of objects in the scene. This predicts a well-known capability of humans that is commonly called *subitization* [10, 30] and it is limited to the first few numbers, a property we observe in our model as well.

Third, the model produces spontaneously a linear structure, which we call *embedding line*, where images are ordered according to quantity. This prediction is strongly reminiscent of the *mental number line* which has been postulated in the psychology literature [31, 32, 33, 34].

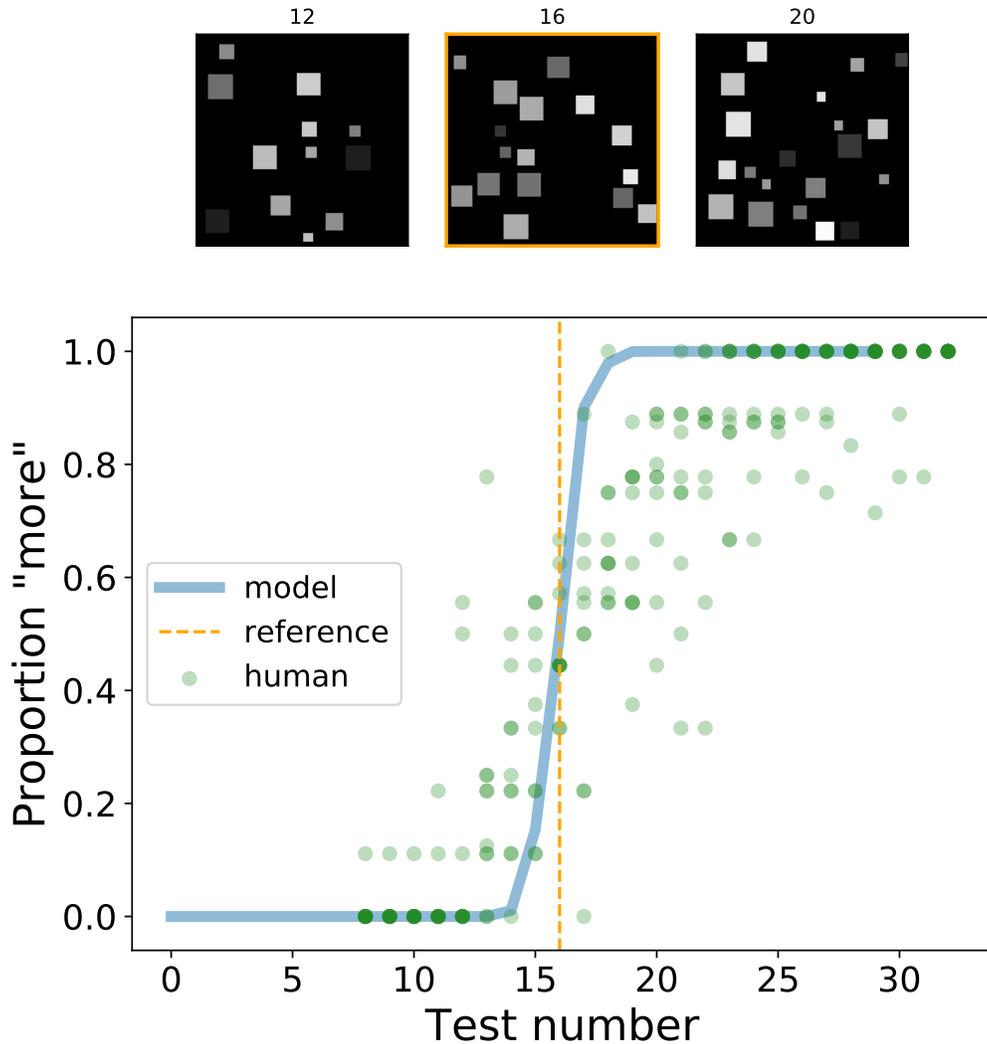


Figure 5: Comparative estimation of quantity. Two images may be compared for quantity [24] by computing their embedding and observing their position along our model’s embedding line: the image that is furthest along the line is predicted to contain more objects. Here images containing a *test number* of objects (see three examples above containing $N=12$, 16 and 20 objects) are compared with images containing the *reference* number of objects (orange line, $N=16$). The number of objects in the test image is plotted along the x axis and the proportion of comparisons that result in a ‘more’ response are plotted on the y-axis (blue line). Human data from 10 subjects [25] is plotted in green.

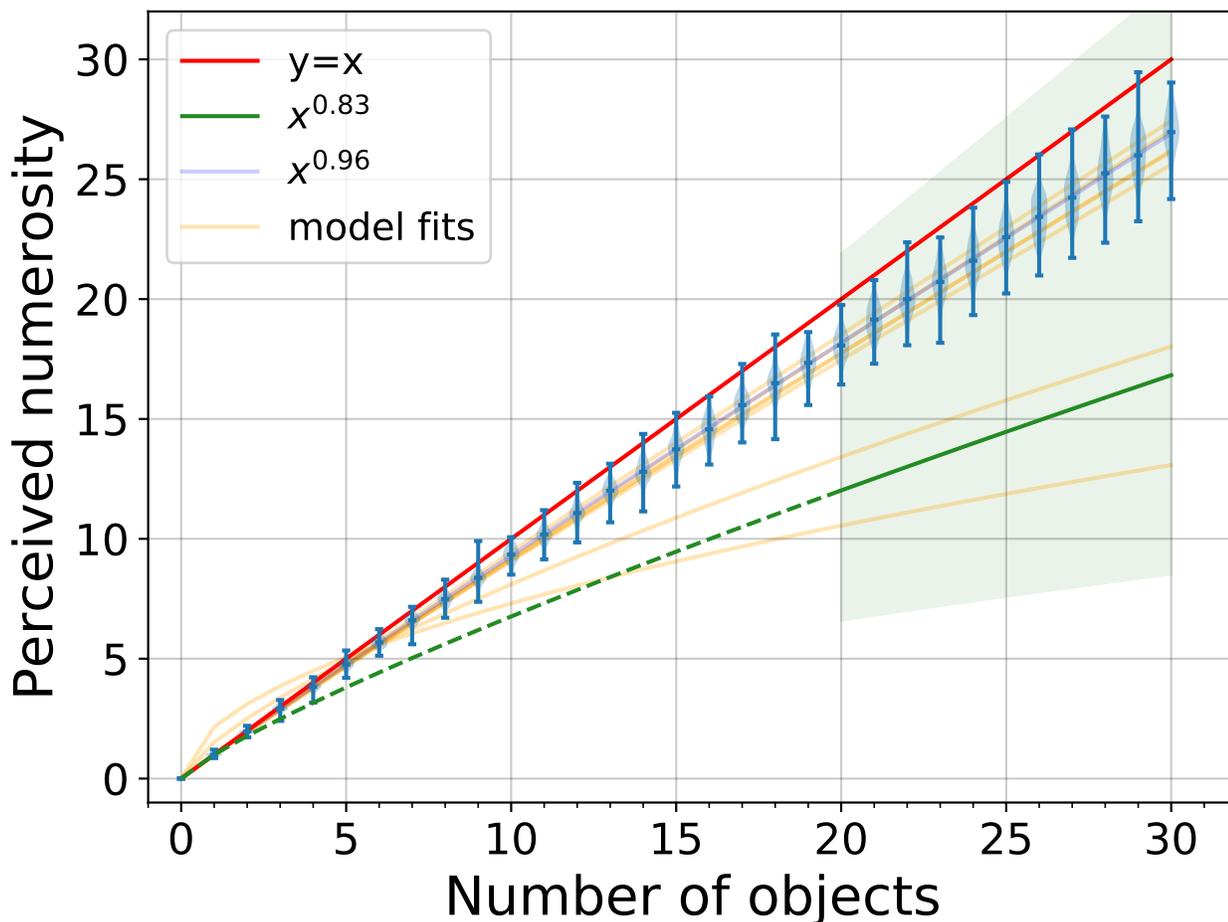


Figure 6: Unsupervised numerosity perception. The position of images in the embedding space fall along a straight line that starts with 0, and continues monotonically with an increasing number of objects. Thus, the position of an image in the embedding line is an estimate for the number of objects in the scene. Here we demonstrate the outputs of such a model, where we rescale the embedding coordinate (an arbitrary unit) so that one unit of distance matches the distance between the “zero” and the “one” clusters. The y-axis represents such perceived numerosity, which is not necessarily an integer value. The red line indicates perfect prediction. Each violin plot (light blue) indicates the distribution of perceived numerosities for a given ground-truth number of objects. The width of the distributions for the higher counts indicates that perception is subject to errors. There is a slight underestimation bias for higher numbers, consistent with that seen in humans [26, 27]. In fact, Krueger shows that human numerosity judgements (on images with 20 to 400 objects) follow a power function with an exponent of 0.83 ± 0.2 . The green line and its shadow depict the range of human numerosity predictions on the same task. The orange lines are power function fits for seven models trained in the same fashion as Model 2 with different random initializations.

The embedding line confers to the model the ability to estimate quantities both in relative comparisons and in absolute judgments. The model predicts the ability to carry out relative estimation, absolute estimation, as well as the tendency to underestimation in absolute judgments. These predictions are confirmed in the psychophysics literature [24, 26].

There is a debate in the literature on whether estimation and subitization are supported by the same mechanisms or separate ones [24, 35]. Our model suggests a solution that will please both sides: both perceptions rely on a common representation, the embedding. However, the two depend on different mechanisms that take input from this common representation. Furthermore, our model predicts that adaptation affects estimation, but not subitization. This is because subitization solely relies on classifiers, which allows for a direct estimate of quantity. Estimation, however, relies on an analog variable, the coordinate along the embedding line, which requires calibration. These predictions are confirmed in the psychophysics literature [26, 24]. Our model predicts the existence of summation units, which have been documented in the physiology literature [22] and have been postulated in previous models [36]. It does not rule out the simultaneous presence of other codes, such as population codes or labeled-line codes [37].

It is important to recognize the limitations of our model: it is designed to explore the minimal conditions that are required for the emergence of a number sense, and abstracts over the details of a specific implementation in the brain. For instance, we limit the model to vision, while it is known that multiple sensory systems may contribute, including hearing and touch [38, 39]. Furthermore, the visual system serves multiple tasks, such as face processing, object recognition, and navigation. Thus, it is likely that multiple visual maps are simultaneously learned, and it is possible that our ‘latent representation’ is shared with other visual modalities [9]. Additionally, we postulate that visually-guided manipulation, and hence the ability to detect and locate objects, is learned before numbers. Thus, it would perhaps be more realistic to consider input from an intermediate map where objects have been already detected and located, and are thus represented as ‘tokens’, in visual space, and this would likely make the model’s task easier, perhaps closer to Model 1 than to Model 2. Making this assumption, however, is not necessary for our observations.

Our investigation adds a concrete case study to the discussion on how abstraction may be learned without explicit supervision. While images containing, say, five objects may look very different from each other, our model discovers a common property, i.e. the number of items, which is not immediately available from the brightness distribution. The mechanism driving such abstraction may be interpreted as an implicit *contrastive learning* signal [40], where the *shake* action identifies pairs of images that ought to be considered as similar, while the *put* and *take* actions signal pairs of images that ought to be considered dissimilar, hence the clustering. However, there is a crucial difference between our model and traditional contrastive learning. In contrastive learning it is the designer who hand-crafts the similarity and dissimilarity training signals in order to achieve an intended learning goal. I.e. the abstraction is directly built into, not discovered by, the network. By contrast, in our model it is the network itself that associates a meaning of ‘close’ and ‘far’ to the P,T,S actions, and ultimately discovers the abstraction. This abstraction is surprisingly strong – while the

primary supervised task, action classification, does not generalize beyond the training limit of three objects, the abstractions of number and quantity extend far beyond it.

Methods

Network Structure

The network we train is a standard deep network [21] composed of two stages. First, a feature extraction network maps the original image of the scene into an embedding space (**Fig. 1A**). Second, a classification network takes the embedding of two sequential images and predicts the action that modified the first into the second (**Fig. 1B**). Given the fact that the classification network takes the embedding of two distinct images as its input, each computed by identical copies of the feature extraction network, the latter is trained in a Siamese configuration [13].

The feature extraction network is a 9-layer CNN followed by two fully connected layers (details in **Fig. S1A**). The first 3 layers of the feature extraction network are from AlexNet [20] pre-trained on ImageNet [41] and are not updated during training. The remaining four convolutional layers and two fully connected layers are trained in our action prediction task.

The dimension of the output of the final layer is a free parameter (it corresponds to the number of features and to the dimension of the embedding space). In a control experiment we varied this dimension from one to 256, and found little difference in the action classification error rates (**Fig. S3**). We settled for a two-dimensional output for the experiments that are reported here.

The classification network is a two-layer fully connected network that outputs a three-dimensional one-hot-encoding vector indicating a put, take or shake action (details in **Fig. S1B**).

Training procedure

The network was trained with a negative log-likelihood loss (NLL loss) function with a learning rate of $1e-4$. The NLL loss calculates error as the $-\log$ of the probability of the correct class. Thus, if the probability of the correct class is low (near 0), the error is higher. The network was trained for 30 epochs with 30 mini-batches in each epoch. Each mini-batch was created from a sequence of 180 actions, resulting in 180 image pairs. Thus, the network saw a total of 162,000 unique pairs of images over the course of training.

We tested for reproducibility by training Model 2 thirty times with different random initializations of the network and different random seeds in our dataset generation algorithm. The embeddings for these reproduced models are shown in **Figure S6**.

Training sets

We carried out separate experiments using synthetic image sequences where objects were represented by randomly positioned squares. The images are 244x244 pixels (px) in size. Objects were positioned with uniform probability in the image, with the exception that they were not allowed to overlap and a margin of at least 3px clearance between them was imposed. We used two different statistics of object appearance: identical size (15px) and contrast (100%) in the first, and variable size (10px - 30px) and contrast (9.8% - 100%) in the second (**Fig. 2**). Mean image intensity statistics for the two training sets are shown in **Figure S2**. The mean image intensity is highly correlated with the number of objects in the first dataset, while it is ambiguous in the second.

Each training sequence was generated starting from zero objects, and then selecting a random action (put, take, shake) to generate the next image. The take action is meaningless on the zero-objects scene and was thus not used there. We also discarded put actions when the objects reached a maximum number. This limit was three for most experiments, but limits of five and eight objects were also explored (**Fig. S5**).

Test sets

In different experiments we allowed up to eight objects per image (**Figs. 3, S5**) and thirty objects per image (**Figs. 4, 5, 6**) in order to assess whether the network can generalize to tasks on scenes containing previously unseen numbers of objects. The first test set was generated following the same recipe as the training set. The second test set was generated to have random images with the specified number of objects (without using actions), this test set is guaranteed to be balanced.

Action classification performance

To visualize how well the model was able to perform the action classification task, we predict actions between pairs of images in our first test set. The error, calculated by comparing the ground truth actions to the predicted actions, is plotted with respect to the number of objects in the visual scene at x_t . 95% Bayesian confidence intervals with a uniform prior were computed for each data point, and a lower bound on the number of samples is provided in the figure captions (**Figs. 3, S3, S5**).

Interpreting the embedding space

We first explored the structure of the embedding space by visualizing the image embeddings in two dimensions. The points, each one of which corresponds to one image, are not scattered across the embedding. Rather, they are organized into a structure that exhibits five salient features: (a) the images are arranged along a one-dimensional structure, (b) the ordering of the points along the line is (almost) monotonic with respect to the number of objects in the corresponding images, (c) images are separated into groups at one end of the embedding,

and these groups are discovered by unsupervised learning, (d) these first few clusters are in one-to-one correspondence with the first few natural numbers, (e) there is a limit to how many number-specific clusters are discovered (**Fig. 4**).

To verify that the clusters can be recovered by unsupervised learning we applied a standard clustering algorithm, and found almost perfect correspondence between the clusters and the first few natural numbers (**Fig. 4B**). The clustering algorithm used was the default Python implementation of HDBSCAN [42]. HDBSCAN is a hierarchical, density based clustering algorithm, and we used the euclidean distance as an underlying metric [43]. HDBSCAN has one main free parameter, the minimum cluster size, which was set to 30 in **Figure 4**. All other free parameters were left at their default values. Varying the minimum cluster size between 5 and 55 does not have an effect on the first few clusters, although it does create variation in the number and size of the later clusters.

One additional structure is not evident from the the embedding and may be recovered from the action classifier: the connections between pairs of clusters. For any pair of images that are related by a manipulation, two computations will be simultaneously carried out; first, the supervised action classifier in the model will classify the action as either P, T, or S (**Fig. 3**) and, at the same time, the unsupervised subitization classifier (**Fig. S4B**) will assign each image in the pair to the corresponding number-specific cluster. As a result, each pair of images that is related by a P action provides a directed link between a pair of clusters (**Fig. S4B, red arrows**), and following such links one may traverse the sequence of numbers in an ascending order. The T actions provide the same ordering in reverse (blue arrows). Thus, the clusters corresponding to the first few natural numbers are strung together like the beads in a necklace, providing an unambiguous ordering that starts from zero and proceeds through one, two etc. (**Fig. S4B**). The numbers may be visited both in ascending and descending order. As we pointed out earlier, the same organization may be obtained more simply by recognizing that the clusters are spontaneously arranged along a line, which also supports the natural ordering of the numbers [44, 45, 33]. However, the connection between the order of the number concepts, and the actions of put and take, will support counting, sum and subtraction.

To estimate whether the embedding structure is approximately one-dimensional and linear in higher dimensions we computed the one-dimensional linear approximation to the embedding line, and measured the average distortion of using such approximation for representing the points. More in detail, we first defined a mean-centered embedding matrix with M points and N dimensions, each point corresponding to the embedding of an image. We then computed the best rank 1 approximation to the data matrix by computing its singular value decomposition (SVD) and zeroing all the singular values beyond the first one. If the embedding is near linear, this rank 1 approximation should be quite similar to the original matrix. To quantify the difference between the original matrix and the approximation, we calculated the element-wise residual (the Frobenius norm of the difference between the original matrix and the approximation), then computed the ratio of the Frobenius norm of the residual matrix and the Frobenius norm of the original matrix. The nearer the ratio is to 0, the smaller the residual, and the better the rank 1 approximation. We call this ratio the

linear approximation error, we show this error compared to some embeddings in **Figure S6**. We computed the embedding for dimensions 8, 16, 64, and 256, (one experiment each) and found ratios of 4.33%, 0.944%, 2.59%, and 1.34%, suggesting that they are close to linear.

Estimating relative quantity

We can use the perceived numerosity to reproduce a common task performed in human psychophysics. Subjects are asked to compare a reference image to a test image and respond in a two-alternative-forced choice paradigm with ‘more’ or ‘less’. We perform the same task using the magnitude of the embedding as the fiducial signal. The model responds with more if the embedding of the test image has a larger perceived numerosity than the reference image. The psychometric curves generated by our model are presented in **Figure 5** and match qualitatively the available psychophysics [24, 27].

Estimating absolute quantity

As noted above, the clusters are spaced regularly along a line and the points in the embedding are ordered by the number of objects in the corresponding images (**Fig. S4**). We postulate that the number of objects in an image is proportional to the distance of that image’s embedding from the embedding of the empty image. Given the linear structure, any one of the embedding features, or their sum, may be used to estimate the position along the embedding line. In order to produce an estimate we use the embedding of the “zero” cluster as the origin. The zero cluster is special, and may be detected as such without supervision, because all its images are identical and thus it collapses to a point. The distance between “zero” and “one”, computed as the pairwise distance between points belonging to the corresponding clusters, provides a natural yardstick. This value, also learned without further supervision, can be used as a unit distance to interpret the signal between 0 and n. This estimate of numerosity is shown in **Figure 6** against the actual number of objects in the image. We draw two conclusions from this plot. First, our unsupervised model allows an estimate of numerosity that is quite accurate, within 10-15% of the actual number of objects. Second, the model produces a systematic underestimate, similar to what is observed psychophysically in human subjects [26].

Acknowledgements

We are very grateful to the Simons Foundation for generously supporting this work. We are in debt to a number of colleagues who provided references to the literature and insightful suggestions: Alessandro Achille, Katie Bouman, David Burr, Eli Cole, Mario Perona, Giovanni Paolini, Markus Meister, Stefano Soatto, Kate Stevenson, Doris Tsao.

References

- [1] Fei Xu, Elizabeth S Spelke, and Sydney Goddard. Number sense in human infants. *Developmental science*, 8(1):88–101, 2005.
- [2] Stanislas Dehaene. *The number sense: How the mind creates mathematics*. OUP USA, 2011.
- [3] Peter Gordon. Numerical cognition without words: Evidence from Amazonia. *Science*, 306(5695):496–499, 2004.
- [4] Pierre Pica, Cathy Lemer, Véronique Izard, and Stanislas Dehaene. Exact and approximate arithmetic in an Amazonian indigene group. *Science*, 306(5695):499–503, 2004.
- [5] Stanislas Dehaene, Elizabeth Spelke, Philippe Pinel, Ruxandra Stanescu, and Sanna Tsivkin. Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science*, 284(5416):970–974, 1999.
- [6] Mengting Fang, Zhenglong Zhou, Sharon Chen, and Jay McClelland. Can a recurrent neural network learn to count things? In *CogSci*, 2018.
- [7] Silvester Sabathiel, James L McClelland, and Trygve Solstad. Emerging representations for counting in a neural network agent interacting with a multimodal environment. In *Artificial Life Conference Proceedings*, pages 736–743. MIT Press, 2020.
- [8] Ivilin Stoianov and Marco Zorzi. Emergence of a ‘visual number sense’ in hierarchical generative models. *Nature neuroscience*, 15(2):194–196, 2012.
- [9] Khaled Nasr, Pooja Viswanathan, and Andreas Nieder. Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Science advances*, 5(5):eaav7903, 2019.
- [10] W Stanley Jevons. The power of numerical discrimination. *Nature*, 3:281–282, 1871.
- [11] Manuela Piazza, Andrea Mechelli, Brian Butterworth, and Cathy J Price. Are subitizing and counting implemented as separate or functionally overlapping processes? *Neuroimage*, 15(2):435–446, 2002.
- [12] Yosef Singer, Yayoi Teramoto, Ben DB Willmore, Jan WH Schnupp, Andrew J King, and Nicol S Harper. Sensory cortex is optimized for prediction of future input. *Elife*, 7:e31557, 2018.
- [13] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säcker, and Roopak Shah. Signature verification using a “siamese” time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.

- [14] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *The Journal of physiology*, 160(1):106, 1962.
- [15] Rüdiger Von der Heydt, Esther Peterhans, and Gunter Baumgartner. Illusory contours and cortical neuron responses. *Science*, 224(4654):1260–1262, 1984.
- [16] Doris Y Tsao, Winrich A Freiwald, Roger BH Tootell, and Margaret S Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–674, 2006.
- [17] Doris Y Tsao, Winrich A Freiwald, Tamara A Knutsen, Joseph B Mandeville, and Roger BH Tootell. Faces and objects in macaque cerebral cortex. *Nature neuroscience*, 6(9):989–995, 2003.
- [18] Chou P Hung, Gabriel Kreiman, Tomaso Poggio, and James J DiCarlo. Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–866, 2005.
- [19] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [21] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [22] Jamie D Roitman, Elizabeth M Brannon, and Michael L Platt. Monotonic coding of numerosity in macaque lateral intraparietal area. *PLoS Biol*, 5(8):e208, 2007.
- [23] Max Wertheimer. *Laws of organization in perceptual forms*. Kegan Paul, Trench, Trubner & Company, 1938.
- [24] David Burr and John Ross. A visual sense of number. *Current biology*, 18(6):425–428, 2008.
- [25] Paula A Maldonado Moscoso, Guido M Cicchini, Roberto Arrighi, and David C Burr. Adaptation to hand-tapping affects sensory processing of numerosity directly: evidence from reaction times and confidence. *Proceedings of the Royal Society B*, 287(1927):20200801, 2020.
- [26] Véronique Izard and Stanislas Dehaene. Calibrating the mental number line. *Cognition*, 106(3):1221–1247, 2008.

- [27] Lester E Krueger. Single judgments of numerosity. *Perception & Psychophysics*, 31(2):175–182, 1982.
- [28] Lisa Feigenson, Stanislas Dehaene, and Elizabeth Spelke. Core systems of number. *Trends in cognitive sciences*, 8(7):307–314, 2004.
- [29] Stanislas Dehaene. Origins of mathematical intuitions: The case of arithmetic. *Annals of the New York Academy of Sciences*, 1156(1):232–259, 2009.
- [30] David Burr, Giovanni Anobile, and Marco Turi. Adaptation affects both high and low (subitized) numbers under conditions of high attentional load. *Seeing and Perceiving*, 24(2):141–150, 2011.
- [31] Frank Restle. Speed of adding and comparing numbers. *Journal of Experimental Psychology*, 83(2p1):274, 1970.
- [32] Stanislas Dehaene, Serge Bossini, and Pascal Giraux. The mental representation of parity and number magnitude. *Journal of experimental psychology: General*, 122(3):371, 1993.
- [33] Stanislas Dehaene, Nicolas Molko, Laurent Cohen, and Anna J Wilson. Arithmetic and the brain. *Current opinion in neurobiology*, 14(2):218–224, 2004.
- [34] Rosa Rugani, Giorgio Vallortigara, Konstantinos Priftis, and Lucia Regolin. Number-space mapping in the newborn chick resembles humans’ mental number line. *Science*, 347(6221):534–536, 2015.
- [35] Samuel J Cheyette and Steven T Piantadosi. A unified account of numerosity perception. *Nature Human Behaviour*, pages 1–8, 2020.
- [36] Tom Verguts and Wim Fias. Representation of number in animals and humans: A neural model. *Journal of cognitive neuroscience*, 16(9):1493–1504, 2004.
- [37] Andreas Nieder. The neuronal code for number. *Nature Reviews Neuroscience*, 17(6):366, 2016.
- [38] Marie Amalric, Isabelle Degenhien, and Stanislas Dehaene. On the role of visual experience in mathematical development: Evidence from blind mathematicians. *Developmental cognitive neuroscience*, 30:314–323, 2018.
- [39] Virginie Crollen and Olivier Collignon. How visual is the number sense? Insights from the blind. *Neuroscience & Biobehavioral Reviews*, 2020.
- [40] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, volume 2, pages 1735–1742. IEEE, 2006.

- [41] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [42] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.
- [43] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer, 2013.
- [44] Stanislas Dehaene and Laurent Cohen. Towards an anatomical and functional model of number processing. *Mathematical cognition*, 1(1):83–120, 1995.
- [45] Marco Zorzi, Konstantinos Priftis, and Carlo Umiltà. Neglect disrupts the mental number line. *Nature*, 417(6885):138–139, 2002.

Supplemental Figures

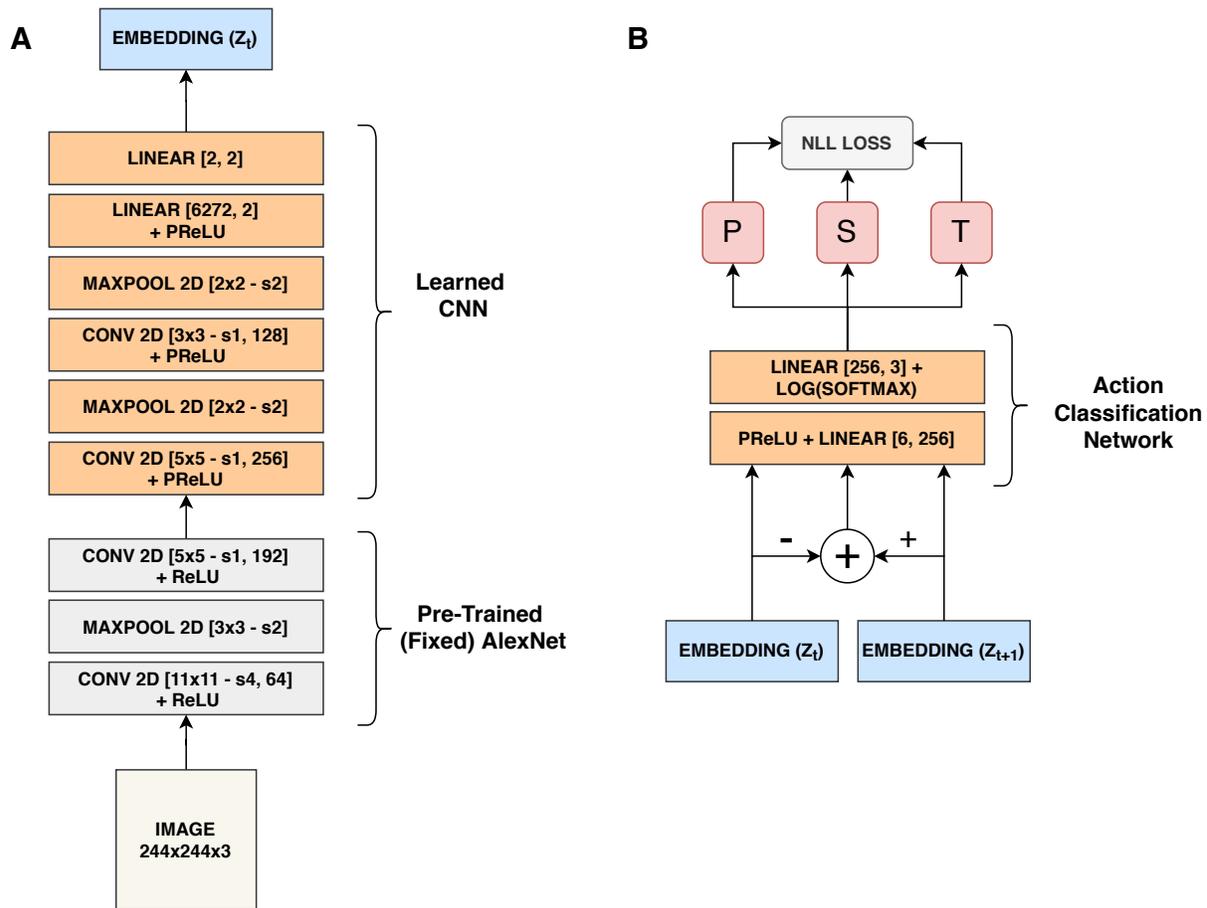


Figure S1: Detailed diagram of the network structure.

(A) The feature extraction / embedding network. The gray layers are pre-trained on ImageNet [41, 20] and remain fixed throughout the course of training. The orange layers are randomly seeded and trained simultaneously with the classifier in (B). The details of the layer are described within the brackets. For example, [11x11 - s4, 64] is an 11x11 kernel with a stride of 4 and 64 filters. During a training step, the embedding network accepts an image (x_t) of the visual scene and generates a lower-dimensional feature embedding (z_t) of the visual scene. An action: (P), (T), or (S) modifies the visual scene and the “after” image (x_{t+1}) is passed through the embedding network as well. The outputs of the embedding network, (z_t) and (z_{t+1}), and their difference ($z_t - z_{t+1}$), are treated as inputs to the action classification network. The shared embedding network is trained together with the classifier (B), in a Siamese configuration. (B) The action classification network is a 2-layer classifier network and is composed of two fully connected layers with a log-softmax transformation on the output. The input is the representation of the visual scene before and after an action is performed. The negative log-likelihood (NLL) loss function is used to train both the action classification network and the embedding network simultaneously. An overview of the entire training paradigm is shown in **Figure 1**.

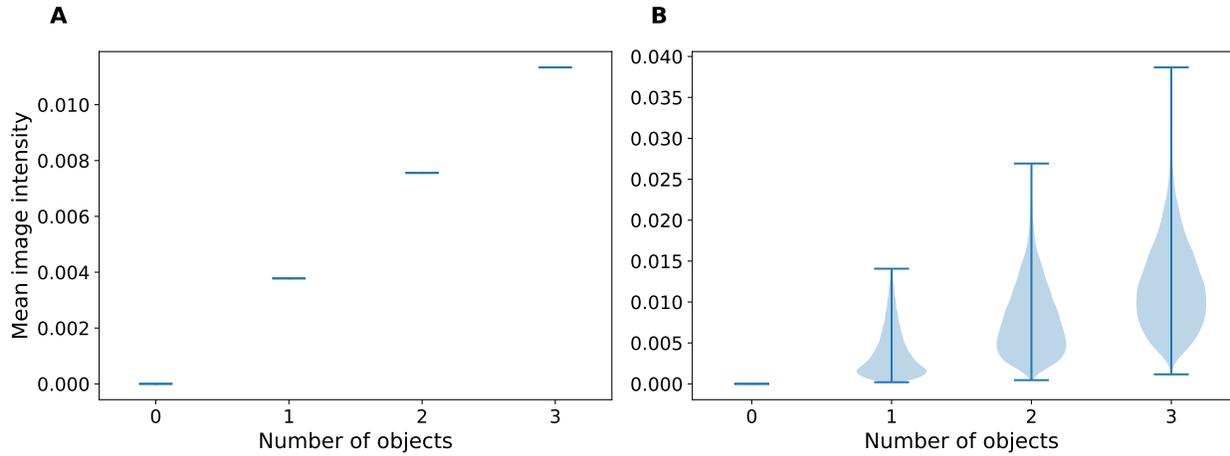


Figure S2: Training set statistics. (A) In dataset 1 (Fig. 2A) objects have the same size and contrast. Thus, the number of objects predicts the mean image intensity and vice-versa. (B) Objects in dataset 2 (Fig. 2B) have variable sizes and variable contrast, thus mean image intensity is not sufficient to predict the number of objects.

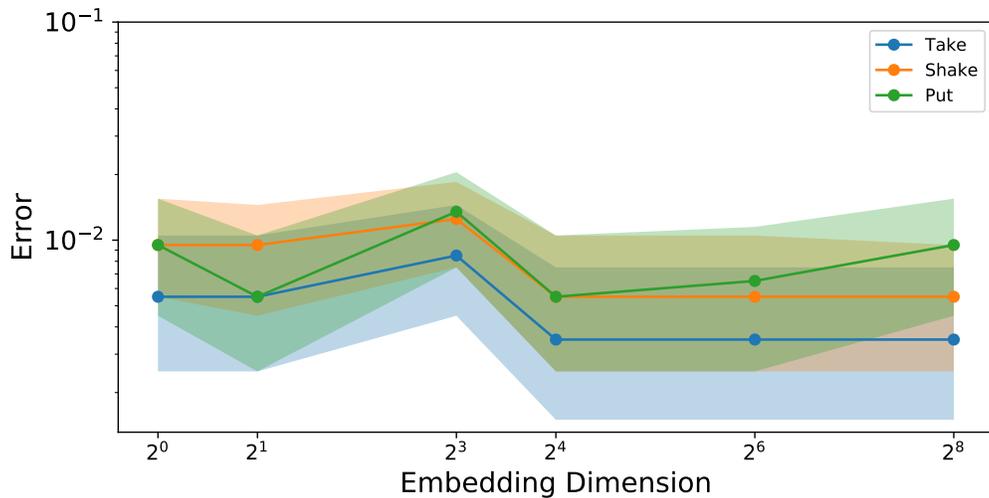


Figure S3: The embedding dimension effects on error. Classification errors for Model 2, averaged over the number of items in the scene (0 - 3) are plotted as a function of the dimension of the embedding (a free parameter in our model). Since the effect is minimal we arbitrarily picked a dimension of two for ease of visualization (Figs. 4B, S4). The shadows show 95% Bayesian confidence intervals ($282 \leq N \leq 355$).

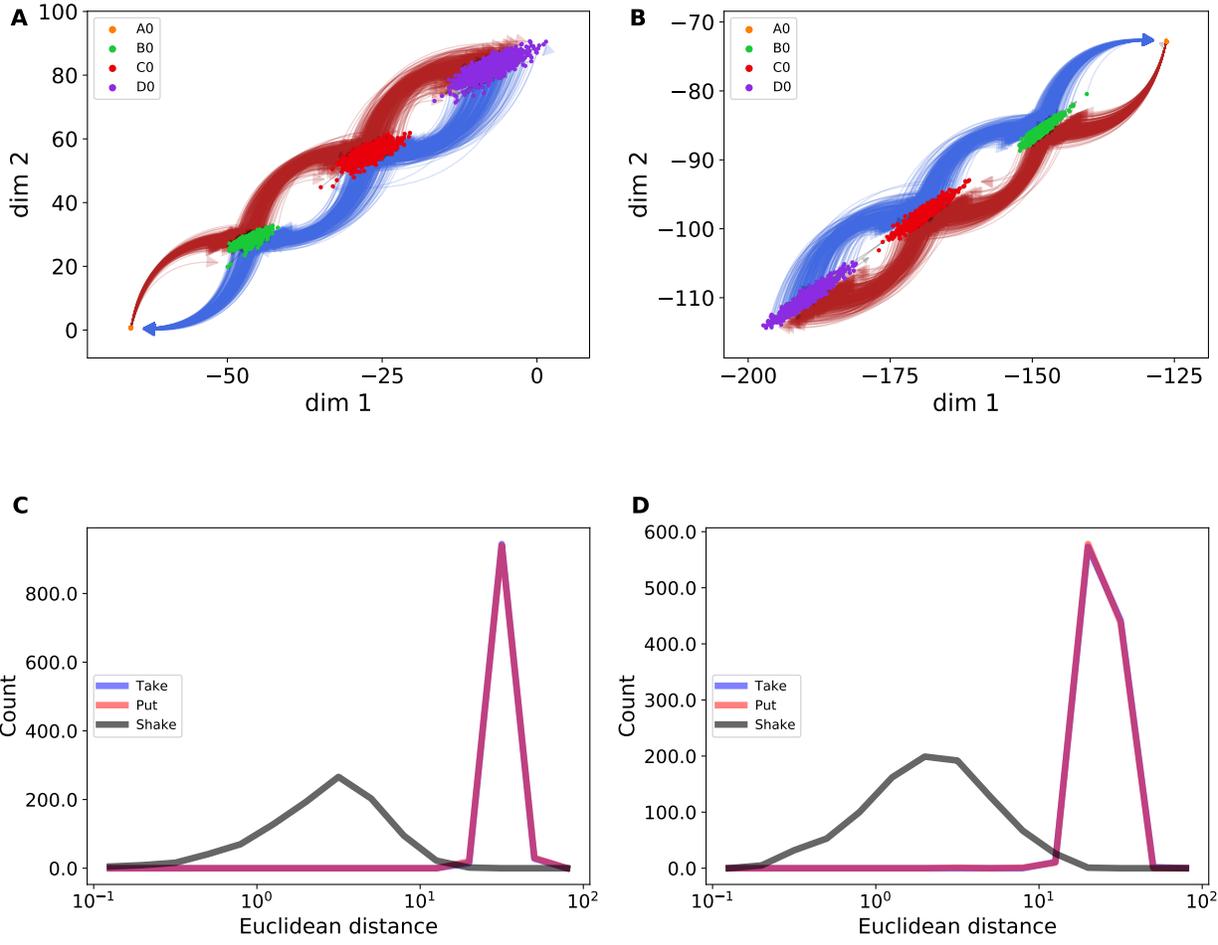


Figure S4: Embeddings with topology for Model 1 and Model 2. A close-up look at the embedding space within the training limit. The left side are plots from Model 1 and the right side from Model 2. **(A)**, **(B)** Unsupervised clustering is performed on the embedding space. Each embedding is colored by its cluster. Each cluster A0 - D0 correspond to images with numerosities 0 - 3. The clusters are well-separated. The “zero” clusters, for both Model 1 and Model 2, are immediately recognizable as they have no variance (orange dot). As numerosity increases, Model 1 clusters remain well-separated, whereas Model 2 clusters begin to come closer to each other. We also overlay a topology from the training actions (P), (T), (S). Blue arrows joining a pair of points represent take actions, red arrows represent put actions. Arrows representing shake actions are under the point clouds and are mostly not visible. **(C)**, **(D)** Distances between pairs of points in the embedding space are histogrammed by action. The histograms show the clearly different distribution for shake actions in comparison to take and put actions. Furthermore, the overlap between shake and non-shake actions is smaller for Model 1 than Model 2, explaining the higher performance in action classification for Model 1.

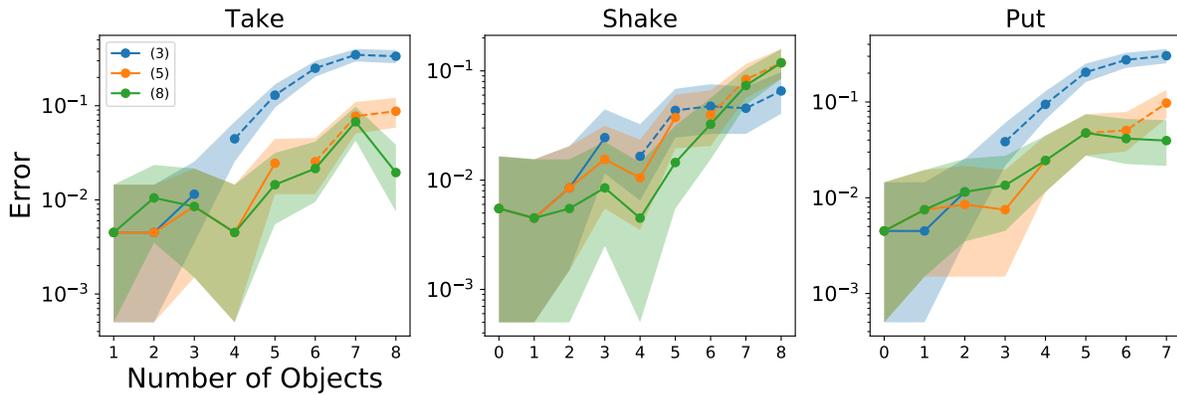


Figure S5: Effect of modifying the training limit. In order to explore the effect of the number of objects during training, we trained the network to predict actions using a maximum of 3, 5, or 8 objects with images like those in dataset 2 (**2B**). We tested the network on 8 objects. Each panels shows errors on the training task and are in the same style as **Figure 3**. The line-breaks and dashed lines mark where the training limit ends and the testing region begins, and the legend shows the training limit in parentheses. The shadows provide 95% confidence intervals ($267 \leq N \leq 355$). As expected, the error is lower when the training limit is higher.

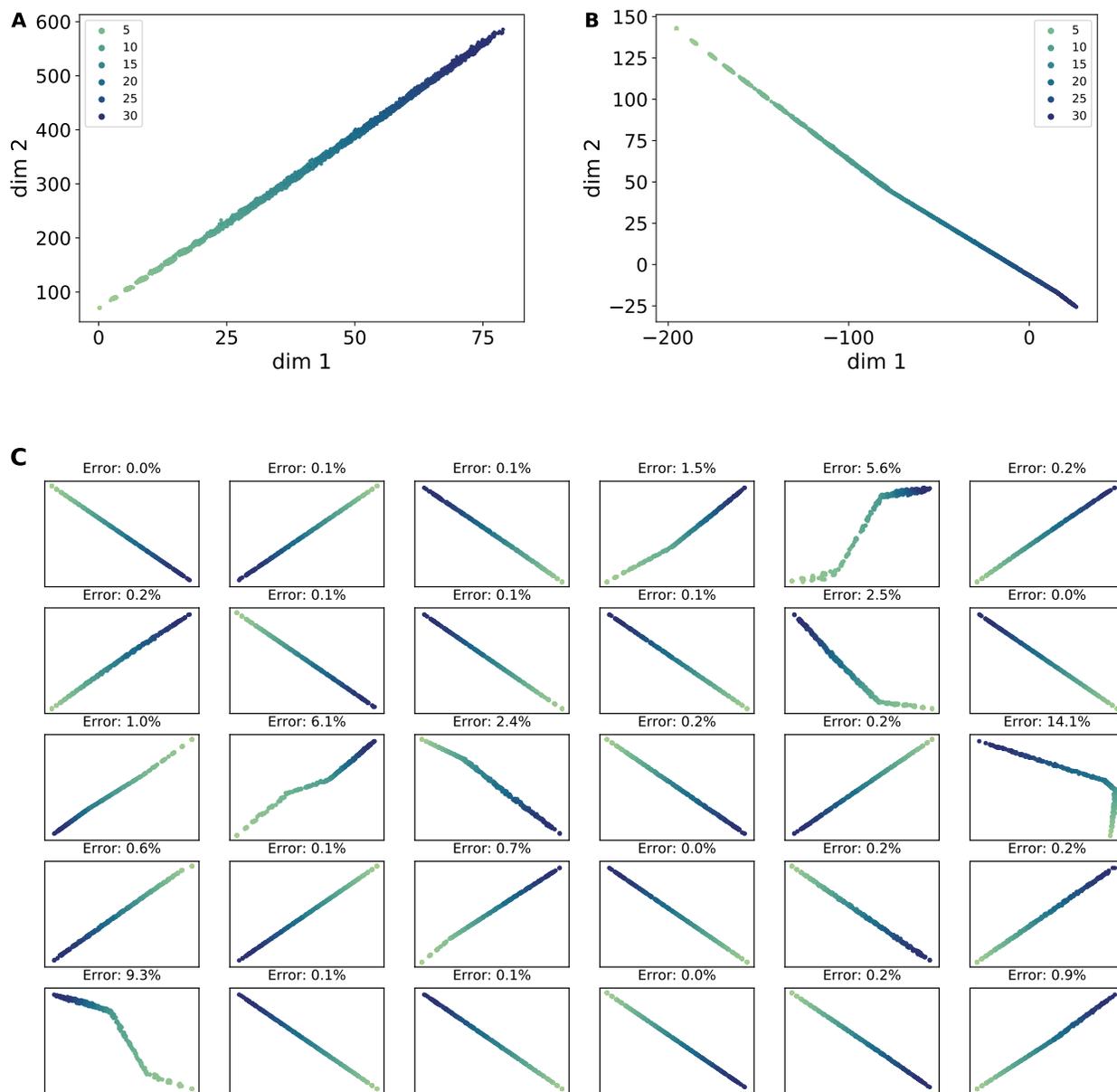


Figure S6: Miscellaneous embedding spaces (A) Embedding space for the network trained on dataset 2, with up to five objects. (B) Embedding space for the network trained on dataset 2, with up to eight objects. (C) Embedding spaces for 30 different random initializations. We repeated the training procedure 30 times on different random initializations of dataset 2, with a training limit of 3 objects. Qualitatively, 21 embedding spaces look like a straight line, five initializations present a slight kink in the line, and four instances either present a large kink or two kinks. The linear approximation error (Methods - Interpreting the Embedding Space) is provided above each subplot and measures the approximate deviation from a purely linear model. An error below 5% predicts an approximately linear embedding line.