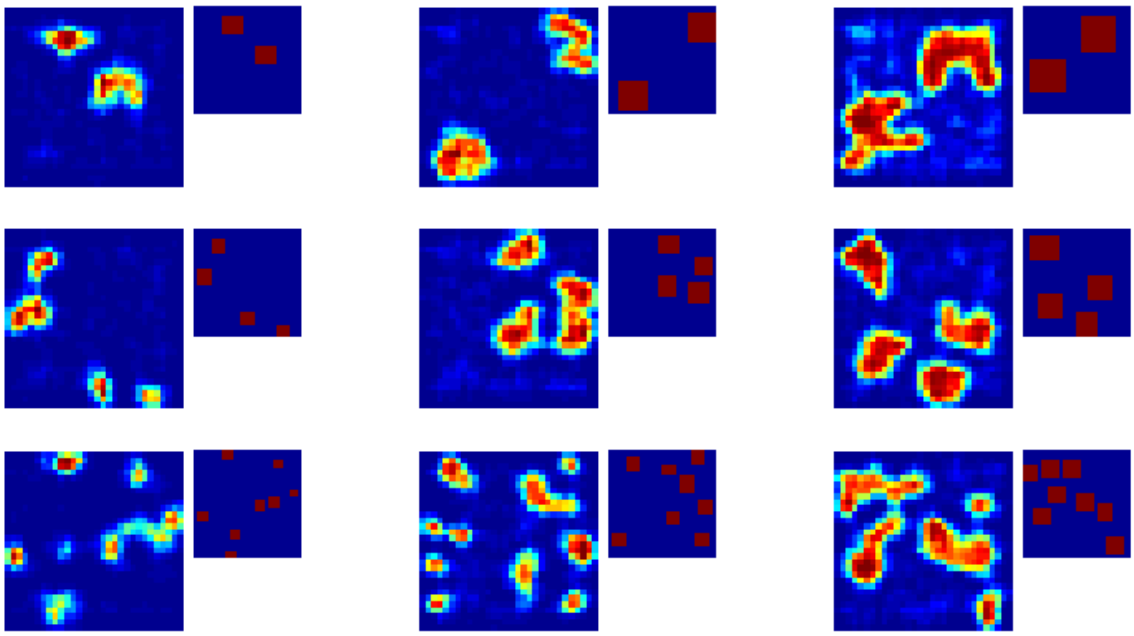
Supplementary Information

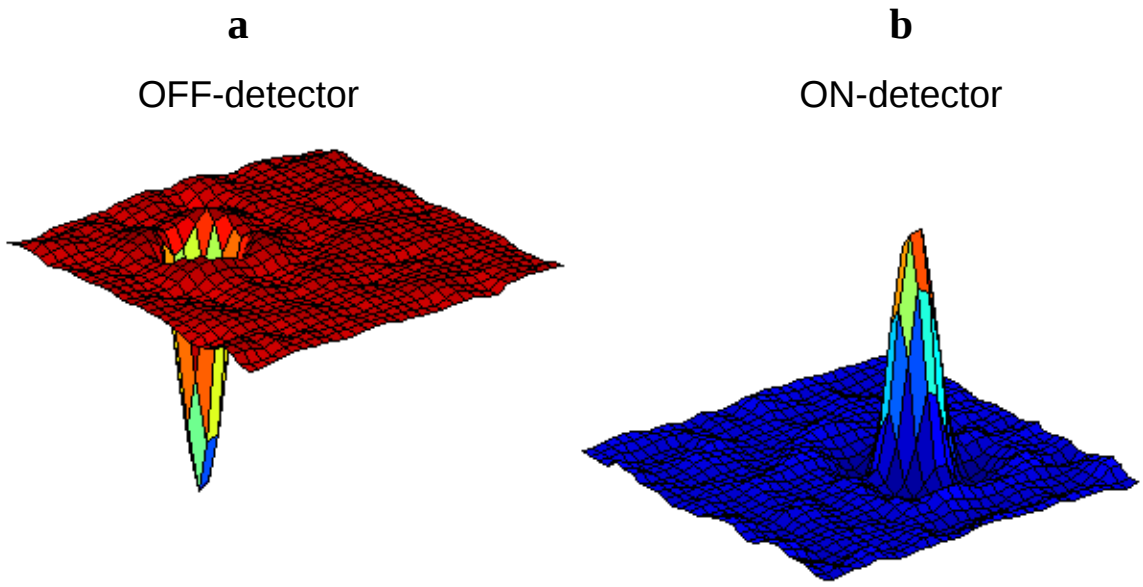
Emergence of a "Visual Number Sense" in Hierarchical Generative Models

Ivilin Stoianov and Marco Zorzi

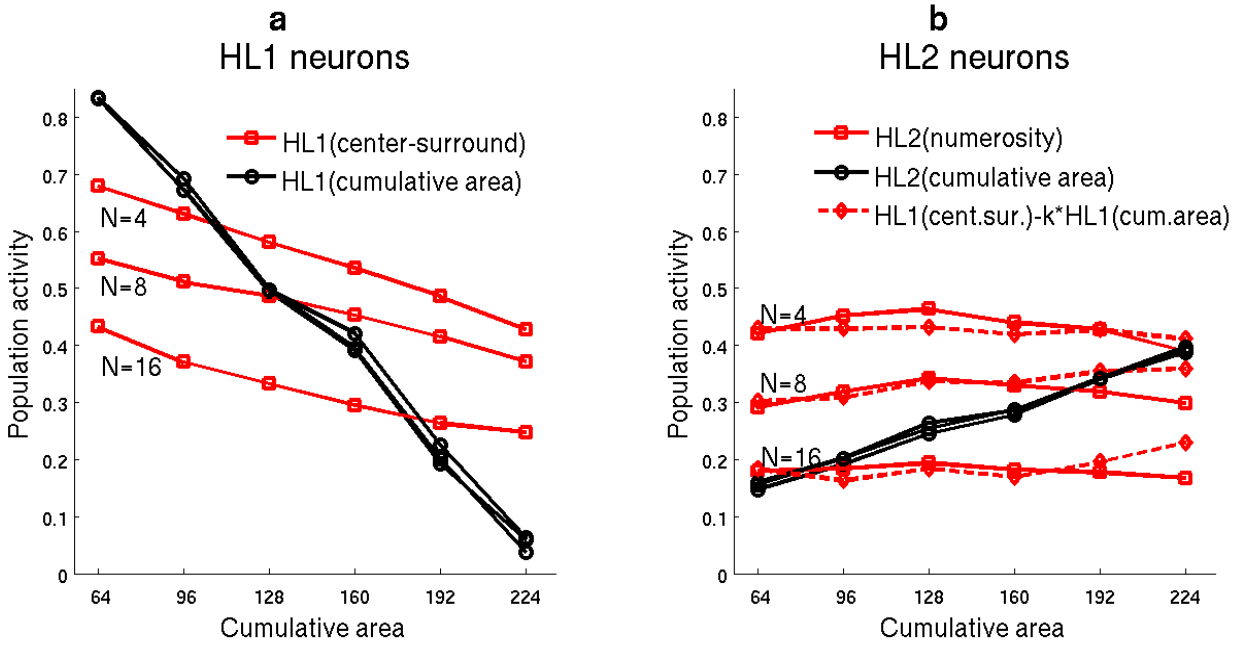
Dipartimento di Psicologia Generale, Università di Padova, Italy



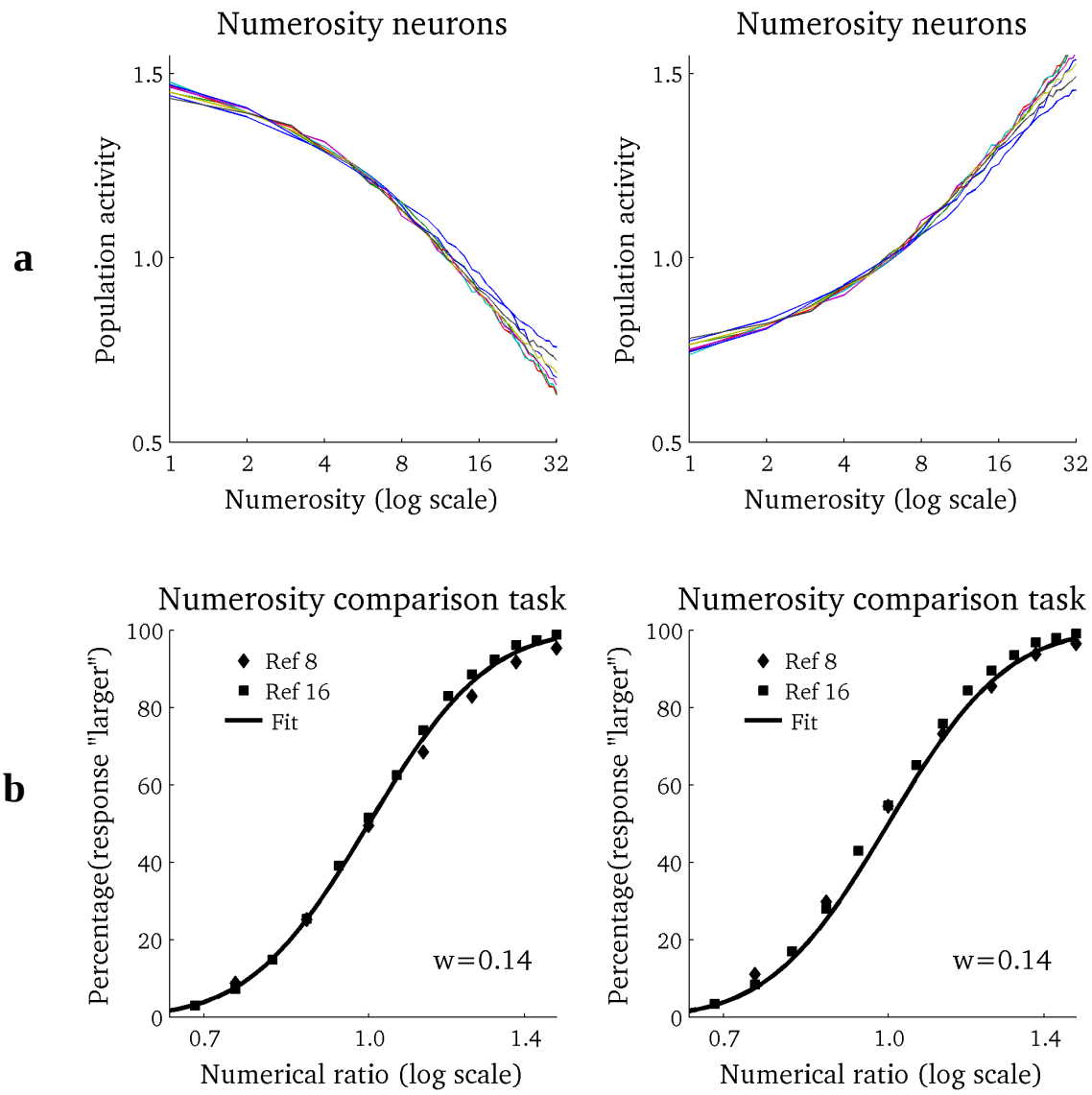
Supplementary Figure 1 Images generated by the deep network at the end of the learning phase when HL2 activity is fed back through the generative weights. The original image is shown to the right of each generated image. Cumulative area increase from left to right and numerosity increases from top to bottom.



Supplementary Figure 2 Response of center-surround neurons emerged in HL1. (**a**) Sample off-center detector and (**b**) sample on-center detector.



Supplementary Figure 3 Numerosity extraction in the deep network. (**a)** Population activity of HL1 center-surround neurons (red lines with squares) is modulated by both numerosity (separate lines for N=4, 8, and 16) and cumulative area (x-axis). The remaining HL1 neurons encode cumulative area invariant to numerosity (black lines with circles). (**b)** The population activity of HL2 numerosity detectors (red lines with squares) is invariant to cumulative-area and can be approximated by a linear combination of the activity of the two types of HL1 neurons (dashed red lines with diamonds). Other HL2 neurons encode cumulative area invariant to numerosity (black lines with circles).



Supplementary Figure 4 Simplified mathematical model. (**a**) Average response of numerosity detectors based on off-detectors (left) and on-detectors (right) for each level of numerosity (abscissa, log-scale) and cumulative surface area (separated lines). (**b**) Performance of the linear classifiers trained to decide whether the visual numerosity was larger than 8 (diamonds) or 16 (squares). Input was the activity of numerosity detectors based on on-detectors (left) or off-detectors (right). Both classifiers yielded a Weber fraction of w=0.14 (fit: sigmoid).

Supplementary Methods

1. Training Database. Each example was a binary 30-by-30 pixel image containing from 1 to 32 randomly-placed non-overlapping rectangular objects, separated by at least 1 pixel. Objects' cumulative surface area noisily ranged from 32 to 256 pixels at 8 levels (with a step of 32). Individual objects were iteratively generated for each image. Object area was initially obtained by dividing the target cumulative area by the number of objects to be generated and adding Gaussian noise (σ=0.15). Two Gaussian random variables (both with σ=0.3) were independently added to the square root of object area (followed by rounding) to define object width and height, respectively. The actual object area was then subtracted from the target cumulative area to start a new iterative step. 200 images were generated for each level of numerosity (n=32) and cumulative surface area (n=8), for a total of 51,200 images. The images were not labeled and learning was only unsupervised.

2. Deep network model. We used a multilayer neural network with one visible layer that represents sensory data and two (hierarchically organized) hidden layers. The layers were connected by both bottom-up (recognition) and top-down (generative) weights. Deep networks can be conveniently reduced to a stack of *Restricted Boltzmann Machines (RBM)*, one for each hidden layer[1]. Each RBM has a layer of feature detectors (hidden units) h_j receiving weighted input $x_j = \sum w_{ji} v_i$ passed through the logistic function $h_j = 1/\left(1+e^{-x_j}\right)$. The first RBM had 80 hidden units and the input was a vector of 900 units encoding vectorized images. Note that the spatial information (e.g., which pixels go together to form one object) is only implicitly coded in the input. The second RBM had 400 hidden units and the input was the activation of the first RBM's hidden layer.

3. Learning. The network was trained to generate the sensory data (i.e., maximizing the likelihood of reconstructing the input data), starting from a given state of the feature detectors

and using the weights w_{ji} in a top-down direction. Contrastive-Divergence learning[2], given an input vector v_i^+, first activates the feature detectors h_j^+ ("positive" phase). Starting from stochastically selected binary states of the feature detectors (using their state h_j^+ as a probability to turn them on), it then infers an input vector v_i^- used in turn to reactivate the features detectors h_j^- ("negative" phase). The weights w_{ji} are updated with a small learning fraction η of the difference between input-output correlations measured in the positive and the negative phases:

$$\delta w_{ij} = \eta\left(v_i^+ h_j^+ - v_i^- h_j^-\right) \quad (1)$$

The two RBM layers were trained in succession, 300 epochs each. Each learning epoch comprised the entire data-set, randomly subdivided into 320 mini-batches. As shown in **Supplementary Fig. 1**, after learning the network was able to generate reasonable reconstructions of the images when activation of the deepest hidden layer was fed back to the visible layer through the top-down weights.

4. Numerosity detectors*:* Sensory magnitudes typically have compressed monotonic neural representations. Such a simple coding of numerosity was indeed found in the LIP monkey cortex[3]. A similar modulation of the BOLD signal was recently found with fMRI in the human homologue of LIP[4]. Thus, we sought detectors of numerosity (*N*) and cumulative surface area (*A*) in the deep network by regressing each neuron's activity h_j with the logarithm of those two properties (all variables normalized) across the entire image database:

$$h_j = \beta_1 \log(N) + \beta_2 \log(A) + \varepsilon \quad (2)$$

Our criterion for a neuron to extract one of those properties was that the regression explained at least 10% of the variance ($R^2 \geq 0.1$) in its activity and the regression coefficient of the complimentary property had an absolute value smaller than 0.10 (the large number of samples made a criterion based on statistical significance too lenient). The "ideal" numerosity detector is invariant to cumulative surface area and thus it is indexed by a large absolute value of the coefficient for numerosity and zero value of the coefficient for cumulative area.

In HL1 we only found feature-detectors extracting c*umulative-surface area* (n=6), with the regressions explaining on average 57% of their response variability (range 40% – 74%). Sensitivity to this feature was also found in many HL2 neurons (n=164; average R^2= 17%; range 10% – 36%). Most importantly, we also found detectors of *numerosity* (n=35) in HL2; the regressions explained on average 22% of their response variability (range 11% – 35%).

5. Network analysis. As noted above, few HL1 neurons computed inversely coded cumulative surface area (n=6). These neurons simply summed activity over the entire stimulus, because their connection weights were roughly uniformly distributed across the 2D input space. All other HL1 neurons can be described as *on-center* (n=21) and *off-center* (n=53) spatial filters (samples in **Supplementary Fig. 2**). The receptive fields of the on- and off-center neurons were uniformly spaced across the 2D input and approximately 6-pixels wide (corresponding to about 1.1 octave-wide 2D-spatial filter). Center-surround neurons are very common in the early visual system (e.g., in LGN)[5].

We further analyzed the 35 HL2 neurons tuned to numerosity (see section 4 above). Notably, each neuron performed localized numerosity estimation, receiving strong input only from HL1 on- and off-center neurons with spatially aligned receptive fields, spanning in total about 10% of the overall input field. The input was positive from off-neurons and negative from on-neurons, thereby providing the same type of local signal. This local input was combined with the negatively weighted activity of HL1 cumulative-area detectors (normalization signal) to yield an output representing the local numerosity. We illustrate this analysis in **Supplementary Fig. 3**. The graphs show the population activity of the different types of HL1 and HL2 neurons as a function of cumulative area (x-axis) and numerosity (three separate lines for numerosities 4, 8 and 16). The activity of center-surround HL1 neurons (population activity of off-detectors minus population activity of on-detectors) was sensitive to both numerosity and cumulative area, whereas HL1 cumulative-area neurons were strongly modulated by area but invariant to numerosity (see panel **a**). In contrast, the activity of the two populations of HL2 neurons identified in Section 4 was sensitive to one dimension and invariant to the other (see panel **b**). Most notably, the population activity of HL2 numerosity neurons, invariant to cumulative area,

could be predicted by a linear combination of the population activity of HL1 center-surround neurons and HL1 cumulative-area neurons (i.e., HL1_center surround – k * HL1_cumulative-area; least square fitting yielded k = 0.29 and R^2=0.98).

6. Numerosity estimation. We used a classic behavioral task, numerosity comparison[6,7], to assess whether the model’s deepest layer (HL2) can support human-like performance. A two-unit linear classifier was trained on the entire training dataset to decide whether a visual numerosity was larger than a reference number (either 8 or 16) by turning on the corresponding yes/no units.

We generated a test dataset that contained the same number of images of the training database (using the same method, see section 1). This was used to test the classifier’s ability to generalize to novel examples and to assess numerosity discrimination. We also generated four smaller control datasets to assess the model’s performance in specific conditions (see refs. 8, 9). Dataset **A** (6,400 images) contained objects with fixed size and shape (3x3 pixels). Therefore, object area was constant but cumulative surface area increased with numerosity. Dataset **B** contained 12,800 images with equal cumulative surface area of 100 pixels. Cumulative area was therefore constant but object area decreased with numerosity. Dataset **C** contained 6,400 images composed of randomly selected objects with variable shape (triangles and ovals), size (object area was 4 or 9 for triangles, 8 or 11 for ovals), and orientation (0 and 90 degrees), for a total of 8 different shapes. This allowed us to assess performance when numerosities were formed by objects with variable features. Dataset **D** contained 25,600 images and had objects spread across the entire image (low density) or confined in an area of 20x20 pixels (high density), all with constant cumulative area of 100 pixels. This allowed us to assess whether performance in numerosity comparison was invariant to density. Since images in the high density condition were markedly different from those in the training database, the classifier was trained on half of dataset D and tested on the other half to perform this test.

To thoroughly assess numerosity discrimination we selected from the test dataset all numerosities that were relatively close to the reference, using a range of numerical ratio (test numerosity / reference) between 0.625 and 1.50 (see ref. 6). For reference 16, this included all

numerosities between 10 and 22, whereas for reference 8 the test numerosities ranged from 5 to 12. On this subset, the classifier accuracy was still 81% correct. The response distributions (probability of “larger” responses) were then used to compute an index of number discriminability, the internal Weber fraction (*w*), which corresponds to the standard deviation of the estimated Gaussian distribution (on a log scale) of the internal representation of numerosity that generates the observed performance (see ref. 6 for details). Thus, the smaller the value of *w*, the better is numerosity discrimination ability, also known as “number acuity”[7,10]. Discrimination ability in humans improves throughout childhood and the mean *w* ranges from about 0.3 in preschoolers to about 0.14 in adults[7,10]. The model’s Weber fraction ($w = 0.15$) compared well against the mean adult value observed in three independent studies ($w = 0.17$[6], 0.15[7], and 0.11[10]). The discriminability index was very similar for the control datasets (from A to D, the *w* values were 0.13, 0.14, 0.14, and 0.17, respectively), thereby confirming the human-like performance of the model in various conditions and invariance to both cumulative surface area and density.

We also performed a series of additional simulations with the deep network to investigate the robustness of numerosity estimation as a function of model parameters. The only parameters that are relatively free regard the number of hidden units (other learning parameters, such as learning rate, were not manipulated and were set to values similar to those used in the seminal paper of Hinton and Salakhutdinov[1] to ensure efficient learning). We found that the model’s Weber fraction improved systematically as a function of HL1 size ($w = 0.26, 0.19, 0.17, 0.15, 0.14, 0.12, 0.13$ for 50, 60, 70, 80, 100, 150, 200 HL1 neurons, respectively; number of HL2 neurons was fixed to 400 as in the main simulation). Greater acuity with more HL1 neurons is not surprising given that center-surround detectors need to cover the input space. In contrast, variation of the Weber fraction was small and not systematic as a function of HL2 size ($w = 0.17, 0.15, 0.15, 0.15$ for 200, 300, 400, 500 HL2 neurons, respectively; number of HL1 neurons was fixed to 80 as in the main simulation).

We also assessed how well the numerosity detectors alone can support numerosity comparison. To this aim, a new linear classifier was trained on the training database, receiving as input the activity of the HL2 numerosity detectors. The generalization performance of the

classifier on the test dataset was still very reasonable: number discriminability was indexed by a Weber fraction of $w = 0.21$. Notably, analysis of the classifier's weights revealed that it performed the comparison task by simply thresholding the weighted sum of the numerosity detectors, thereby corroborating our finding that the population activity of the HL2 numerosity detectors encodes numerosity information (cf. section 5).

7. Control simulations. We also performed a series of control simulations to determine whether numerosity information can be directly extracted from the image data or from the first hidden layer (HL1) of the deep network. First, a linear classifier directly fed with the visual input was unable to learn the numerosity comparison task. Though lacking any plausibility for neuro-cognitive modeling, we also tested a more powerful learning algorithm, Support Vector Machines (SVM)[11]. SVMs with linear kernels, after training on subsets of various sizes (50 or 100 images for each level of numerosity and cumulative surface area), showed very poor generalization performance and fully inadequate numerosity discrimination performance (as shown by a Weber fraction of 0.91 and 0.94, respectively). SVMs with Gaussian kernels could not generalize at all.

In contrast, a linear classifier fed with the activity of the entire HL1 yielded a Weber fraction of $w = 0.24$. Although this value indicates much poorer numerosity discrimination performance (the w is similar to that of young children[6]) compared to the classifier trained on HL2 , this result is consistent with our finding that numerosity information can be extracted by a linear combination of the activities of HL1 neurons (see Section 5). Indeed, when the classifier was prevented from combining the activity of center-surround HL1 neurons and cumulative-area HL1 neurons by restricting the input to either type of information, performance dropped to a fully inadequate level ($w = 0.42$ and $w = 1.19$, respectively). These analyses show that numerosity information is progressively extracted across hidden layers in the deep network.

8. Simplified mathematical model. Building on the analyses of the neural network model, we tested a simpler mathematical model governed by just few parameters. The first layer (size: 13 x 13) consisted of uniformly spread off-detectors O^{ij} :

$$O^{ij} = f\left(\sum V^{ij} I + 1\right) \quad (3)$$

where I is the input image, V^{ij} are 2D-Gaussian-shaped weights ($\sigma = 2$) defining the detector's receptive field (for simplicity, surround suppression was not modeled), and $f(.)$ is the logistic function. The second layer (size: 6 x 6) consisted of uniformly spread local numerosity detectors N^{kl} receiving the activity of layer-1 off-detectors and a normalization signal c :

$$N^{kl} = \sum W^{kl} O + c \quad (4)$$

where W^{kl} are 2D-Gaussian-shaped weights ($\sigma = 10$). The term c represents the image (log) cumulative surface area and is defined as follows:

$$c = \log\left(1 + \frac{\sum I}{c_{max}}\right) \quad (5)$$

where c_{max} is the maximal cumulative surface area across the image data.

The model was assessed in terms of internal coding and capacity to support numerosity comparison using the entire image database. The numerosity detectors showed monotonic coding of numerosity on a log scale, as in the deep network model. Population activity (average across numerosity detectors) decreased as a function of log numerosity (R^2 = 0.83) and was not modulated by cumulative surface area (**Supplementary Fig. 4a**, left). A linear classifier trained to perform numerosity comparison and tested on the novel test set yielded a Weber fraction of *w* = 0.14 (**Supplementary Fig. 4b**, left). Thus, the simplified model faithfully represented the learned model; the greater accuracy of the former in representing numerosity is the consequence of a more regular spatial structure and the absence of learning-derived noise. A control model with no normalization signal performed much worse: log numerosity explained only 46% of variance in the population activity of numerosity detectors and the Weber fraction of the linear classifier increased to *w* = 0.39.

Monotonic coding of numerosity in monkey LIP neurons has been observed in two variants, with firing rates that either increase or decrease as a function of numerosity[3]. We therefore tested

a version of the model where off-detectors were replaced by on-detectors and the normalization signal had opposite sign. In this model, the activation of the numerosity detectors monotonically increased with numerosity. Population activity increased as a function of log numerosity (R^2 =0.83) and was not modulated by cumulative surface area (**Supplementary Fig. 4a**, right). Performance in numerosity comparison was identical to the first version of the model ($w = 0.14$, **Supplementary Fig. 4b**, right). Finally, combining both on- and off-detectors in a single model, whereby numerosity was coded with both monotonically increasing and decreasing activation, did not increase performance in numerosity comparison ($w = 0.14$).

References

1. Hinton, G. & Salakhutdinov R. *Science* **313**, 504 (2006).
2. Hinton, G., Osindero, S. & Teh, Y. *Neural Comput*. **18**, 1527 (2006).
3. Roitman, J., Brannon, E. & Platt, M. *PLoS biology* **5**, e208 (2007), doi: 10.1371/journal.pbio.0050208.12
4. Santens, S. et al. *Cerebral cortex* **20**, 77-88 (2010).
5. Palmer, S. E. *Vision Science*. MIT Press, Cambridge, MA (1999).
6. Piazza, M. et al. *Neuron* **44**, 547 (2004).
7. Piazza M. et al. *Cognition* **116**, 33 (2010).
8. Brannon, E.M. & Terrace, H. S. *Science* **282**, 746 (1998).
9. Nieder, A., Freedman, D. & Miller, E.K. *Science* **297**, 1708-1711(2002).
10. Halberda, J. & Feigenson, L. *Developmental psychology* **44**, 1457-65(2008).
11. Cortes, C. & Vapnik, V. *Machine Learning* **20**, 273-297 (1995).