SUBITIZING WITH VARIATIONAL AUTOENCODERS

Rijnder Wever, Tom Runia — Intelligent Sensory Information Systems

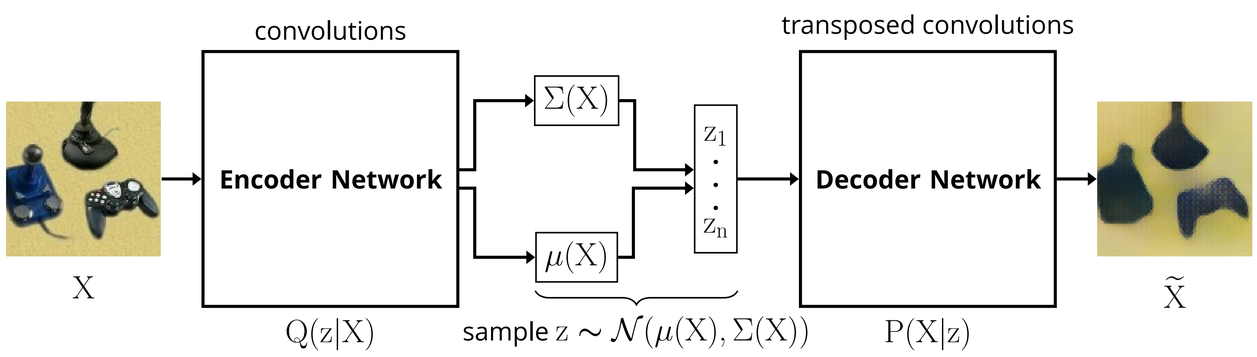
University of Amsterdam

Numerosity, the number of objects in a set, is a basic property of visual scenes. Inspired by Stoianov & Zorzi [1] we propose an **unsupervised** generative model to learn visual numerosity representations from natural and synthetic image datasets designed for instance counting within the subitizing range. Specifically, we use a hierarchically organized **convolutional variational autoencoder** (VAE) for encoding and reconstructing training images. Provided that numerosity is a key characteristic in the images, the network will learn to encode visual numerosity in the latent representation.

CONTRIBUTIONS

- Our VAE spontaneously encodes numerosity in the subitizing range when trained on natural images.
- Numerosity and object area are represented separately, similar to biological neural networks.
- A loss function aimed at contour reconstruction improves the network's subitizing performance.
- Numerosity information represented in the latent space is successfully extrapolated to a subitizing task.

VARIATIONAL AUTOENCODER



VAEs are generative models that perform **unsupervised** representation learning. The VAE's objective function is the combination of a reconstruction term and a KL regularization:

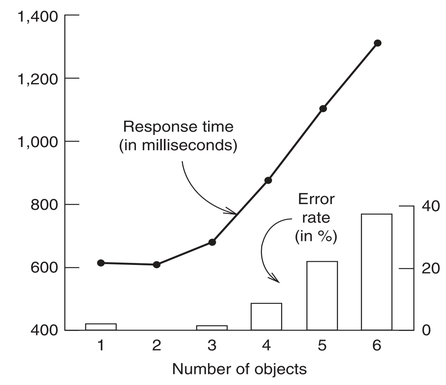
$$\mathcal{L}_{\text{VAE}} = \mathrm{E}[\log \mathrm{P}(\mathrm{X} \mid \mathrm{z})] - \mathcal{D}_{\text{KL}}[\mathrm{Q}(\mathrm{z} \mid \mathrm{X}) \,||\, \mathrm{P}(\mathrm{z})] \tag{1}$$

SUBITIZING

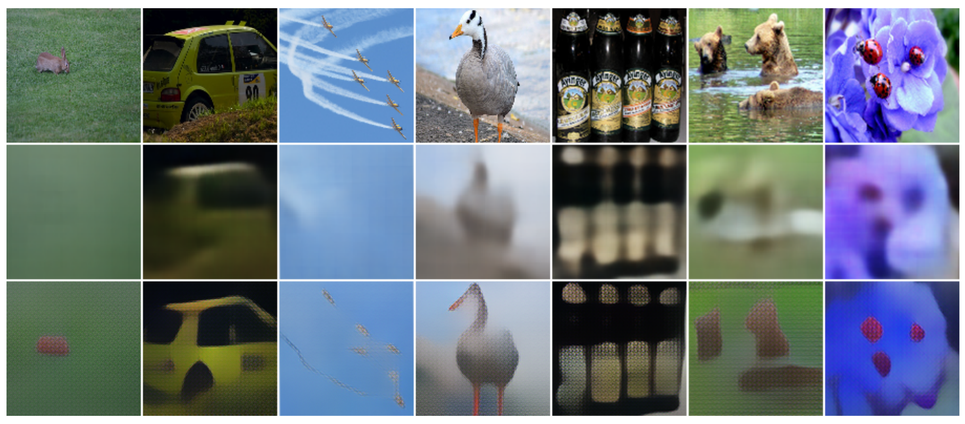
Subitizing is a perceptual ability that enables rapid and spontaneous identification of the numerosity of small visual sets. When the **subitizing range** of 1 – 4 instances is exceeded, other cognitive mechanisms related to instance counting are invoked. We explore the emergence of visual number sense in unsupervised deep networks trained on **natural images** from the **Salient Object Subitizing Dataset** [2].

VISUAL NUMBER SENSE

- **Parallel processing** achieves rapid perceptual discrimination of exact numerosity [3].
- The visiospatial system performs **nonmediated** spatial map generation of salient objects [4].
- Neural populations can automatically develop the ability to judge visual numerosity.



OPTIMIZING THE RECONSTRUCTIONS



Feature Perceptual Loss [5] improved object contour reconstruction by using intermediate layer representations from a pretrained network in the objective function of the VAE.

Augmentation of the SOS dataset with **synthetic data** was required to familiarize the model with an extensive distribution of object types and the spatial configurations thereof.

SIZE-INVARIANT NUMEROSITY DETECTORS

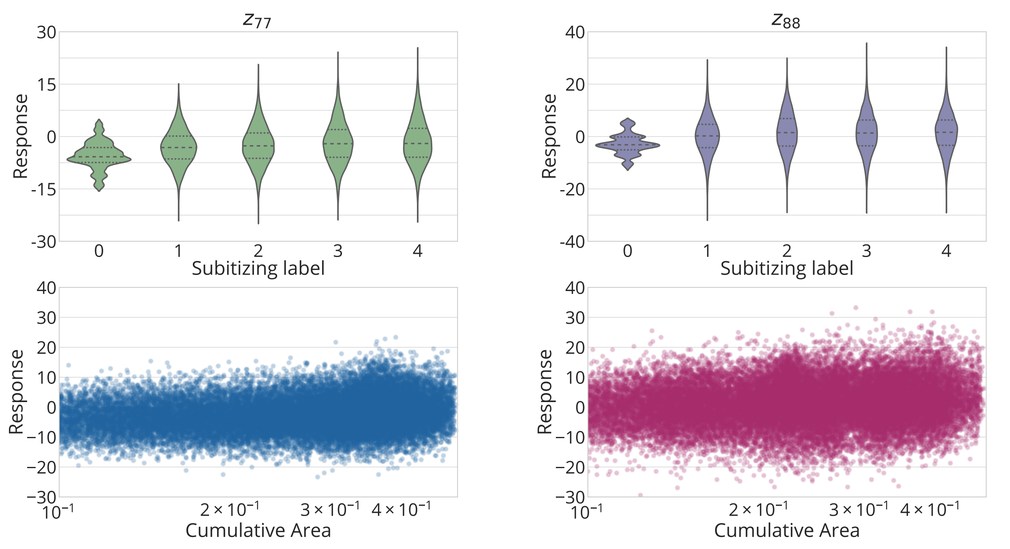
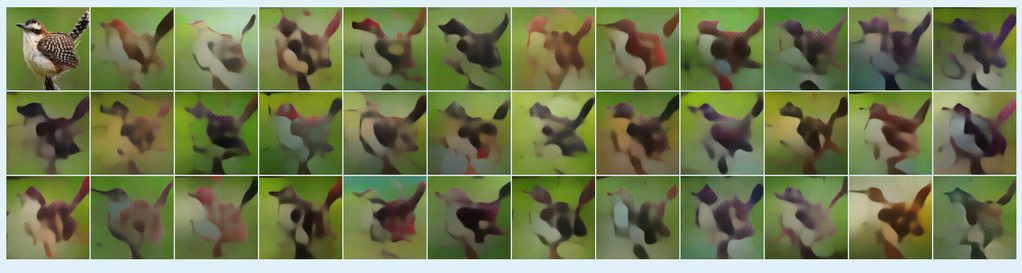


Figure 1. Responses of latent dimensions to changes in object area and numerosity.



Latent dimensions of the learned representations encode complex information. Numerosity is likely encoded **invariant of cumulative object area**, as in biological neural networks.

SUBITIZING TASK PERFORMANCE

Table 1. Average classification precision (%) of count labels to images from the SOS test dataset.

Count Label →	0	1	2	3	4+	mean
Chance	27.5	46.5	18.6	11.7	9.7	22.8
GIST	67.4	65.0	32.3	17.5	24.7	41.4
SIFT+IFV	83.0	68.1	35.1	26.6	38.1	50.1
CNN_FT	93.6	93.8	75.2	58.6	71.6	78.6
VAE + softmax (ours)	76.0	49.0	40.0	27.0	30.0	44.4

We compare our **unsupervised approach** to existing supervised approaches to instance counting. The strength of the representation learned by the VAE is measured by training a simple **softmax classifier** that is fed latent representations of images with corresponding count labels.

[1] Stoianov I. & Zorzi M. Emergence of a 'visual number sense' in hierarchical generative models. *Nature Neuroscience*, 2012.

[2] Zhang J. et al. Salient object subitizing. *IJCV*, 2017.

[3] Dehaene S. The number sense: How the mind creates mathematics. OUP USA, 2011.

[4] Piazza M. & Izard V. How humans count: numerosity and the parietal cortex. *The Neuroscientist*, 2009.

[5] Hou X. et al. Deep feature consistent variational autoencoder. *WACV*, 2017.