SUBITIZING WITH VARIATIONAL AUTOENCODERS

Rijnder Wever, Tom Runia

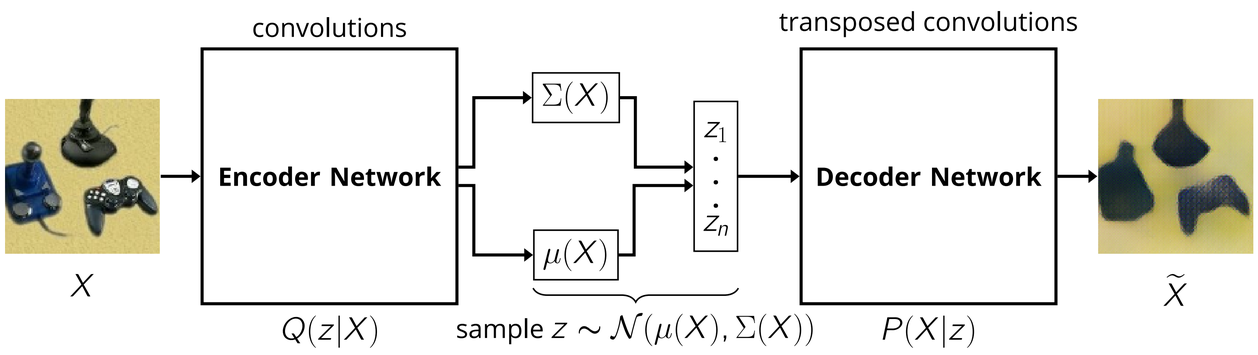
University of Amsterdam

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 catered to instance counting within the subitizing range. Specifically, we employ a hierarchically organized **convolutional variational autoencoder** (VAE) tasked with 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

- Visual numerosity in the subitizing range emerges as a statistical property of natural images in VAEs.
- Numerosity and object area are represented seperately, similair to biological neural networks.
- A loss function aimed at contour reconstruction improves the network's subitizing performance.
- Numerisoty information represented in the latent space is succesfully extrapolated to a subitizing task.

VARIATIONAL AUTOENCODER



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

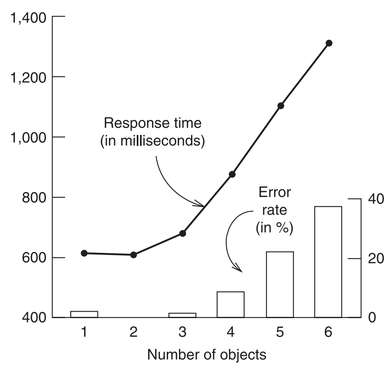
$$\mathcal{L}_{VAE} = E[\log P(X \mid z)] - \mathcal{D}_{KL}[Q(z \mid X) \mid\mid P(z)] \quad (1)$$

SUBITIZING

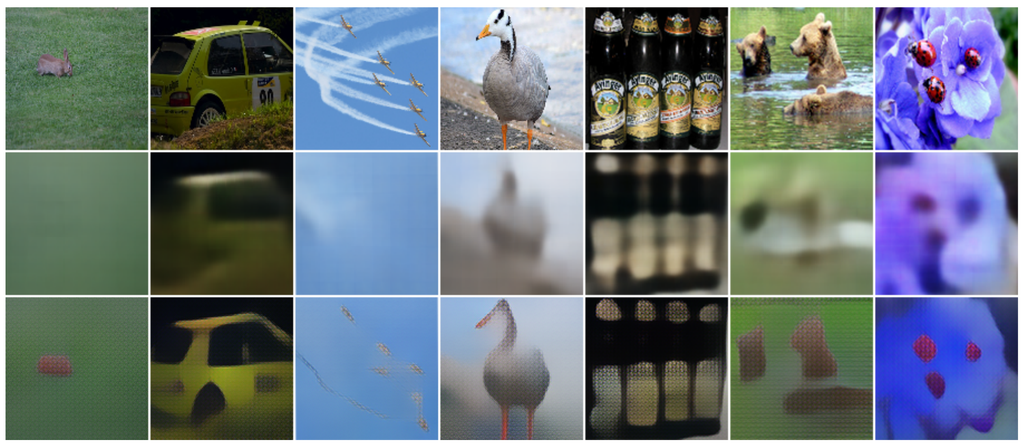
Subitizing is a perceptual ability that enables rapid and spontaneous identification of the numerosity of small visual sets. When the **subitizing range** of 1 – 4 instances is exceeded, other cognitive mechanisms related to instance counting are employed. We explore the emergence of visual number sense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

VISUAL NUMBER SENSE

Immediate but limited exact **visual numerosity** processing abilities could be a result of a capacity-limited multiple object individuation mechanism that performs **parallel processing** of salient objects [3]. The cause of it's **sudden** character is hypothesized to be a byproduct of the visiospatial system's automatic salient object map generation of a limited number of salient locations in the visual field [4].



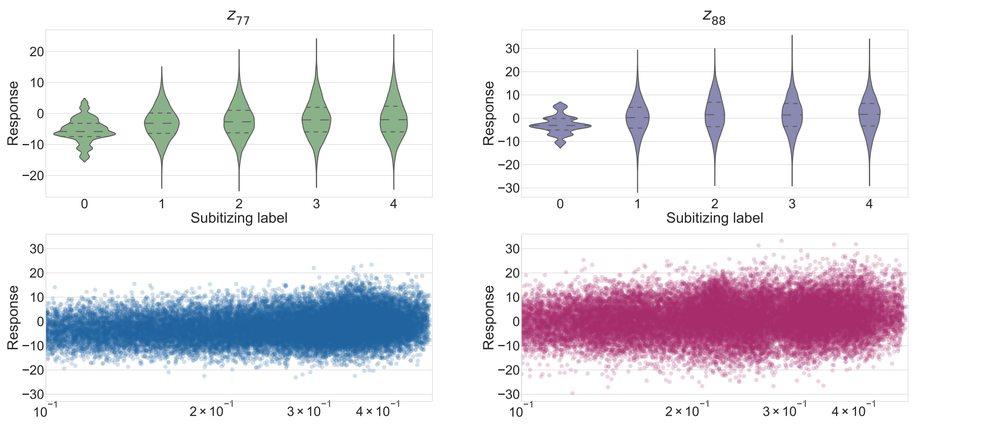
VISUAL OPTIMIZATION RESULTS

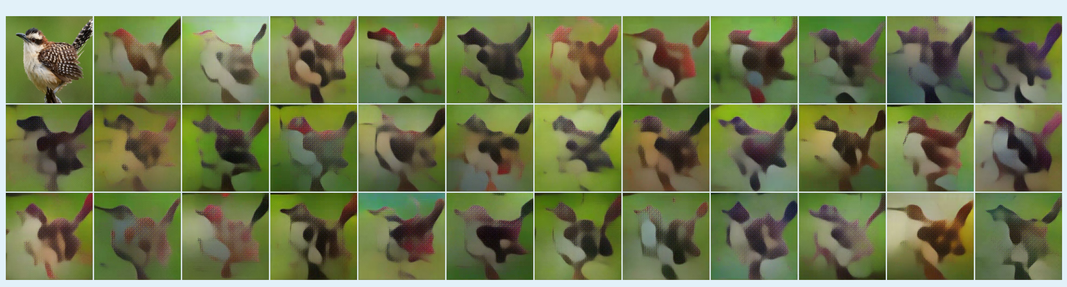


Reconstruction of figure-ground organization was improved by employing **Feature Perceptual Loss** [5] which uses intermediate layer representations from a pretrained network in the objective function of the VAE.

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





Comparable with biological neural networks, an analysis of the learned representations revealed that numerosity is represented **invariant to cumulative object area**. However, no latent dimension represents just one visual property.

SUBITIZING TASK PERFORMANCE

Table: Average classification precision (%) of count labels across various algorithms.

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] Poncet M. et al. Individuation of objects and object parts rely on the same neuronal mechanism. *Scientific reports*, 2016.

[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.