SUBITIZING WITH VARIATIONAL AUTOENCODERS

Rijnder Wever, Tom Runia

University of Amsterdam

Numerosity, the number of objects in a set, is a basic property of a given visual scene. Inspired by Stoianov and 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.

SUBITIZING

Subitizing is a perceptual ability that enables near-instantaneous and spontaneous identification of the numerosity of small visual sets. When the **subitizing range** of 1 – 4 instances is exceeded, other cognitive mechanisms are employed to arrive at an instance count.

We explore the emergence of visual number sense in deep networks trained in an unsupervised setting on **natural images** from the **Salient Object Subitizing Dataset** [2]. The complexity of natural visual scenes helps capture the abstract nature of number sense.

VARIATIONAL AUTOENCODER

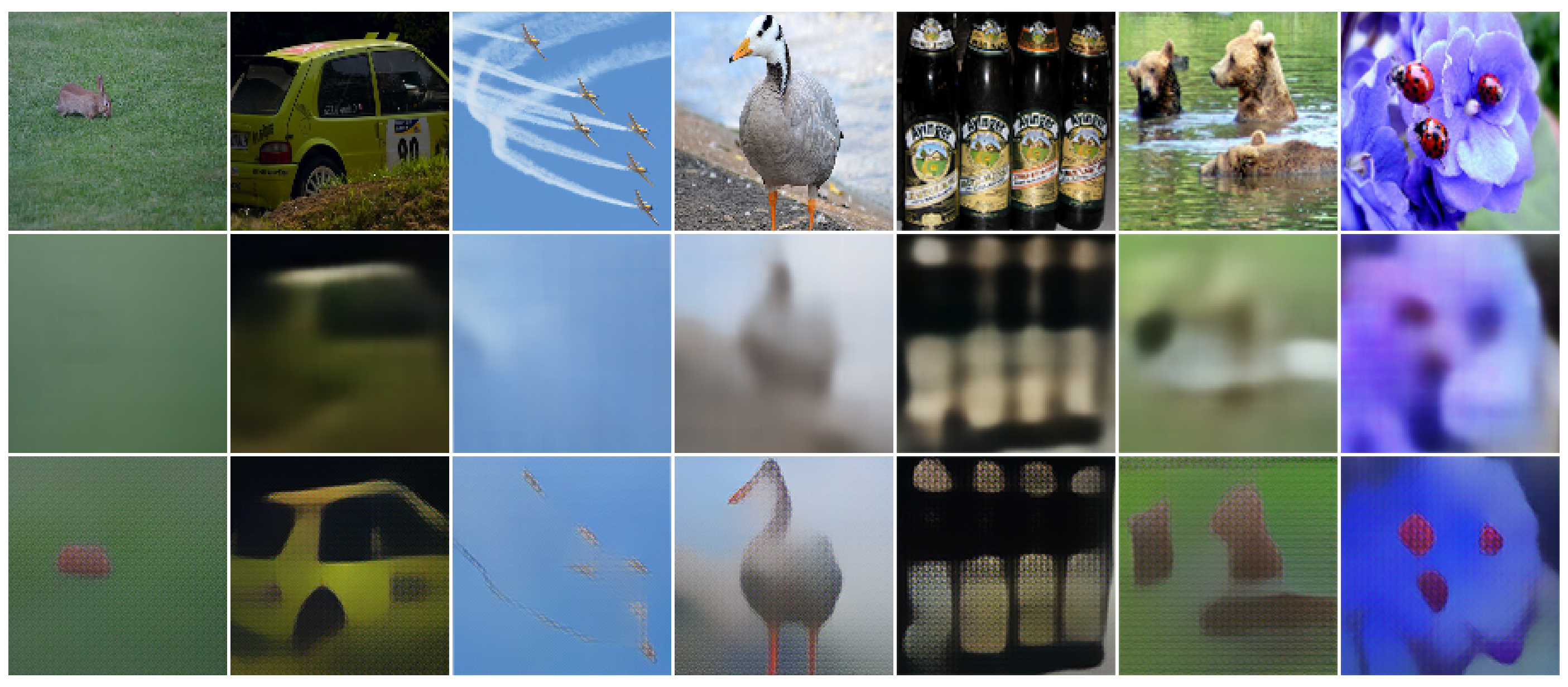
VAEs are generative algorithms that perform **unsupervised** representation learning. Instead of mapping data samples X to a deterministic latent representation as in conventional autoencoders, VAEs learn a posterior distribution $Q(z \mid X)$. Input reconstruction proceeds by sampling latent vectors z from Q and passing them through a decoder network. We parameterize the encoder and decoder network with deep CNNs.

The VAE's objective function is the summation of a reconstruction term and a KL regularization:

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

Sampling is made feasible by parameterizing Q as a Gaussian with learned mean and variance parameters. P(z) is typically a unit Gaussian.

OPTIMIZATIONS



We observed difficulties with reconstructing multiple salient objects, negatively affecting the ability to subitize. Therefore, we employ the recent **Feature Perceptual Loss** [3] which uses intermediate layer representations from a pretrained network in the objective function of the autoencoder.

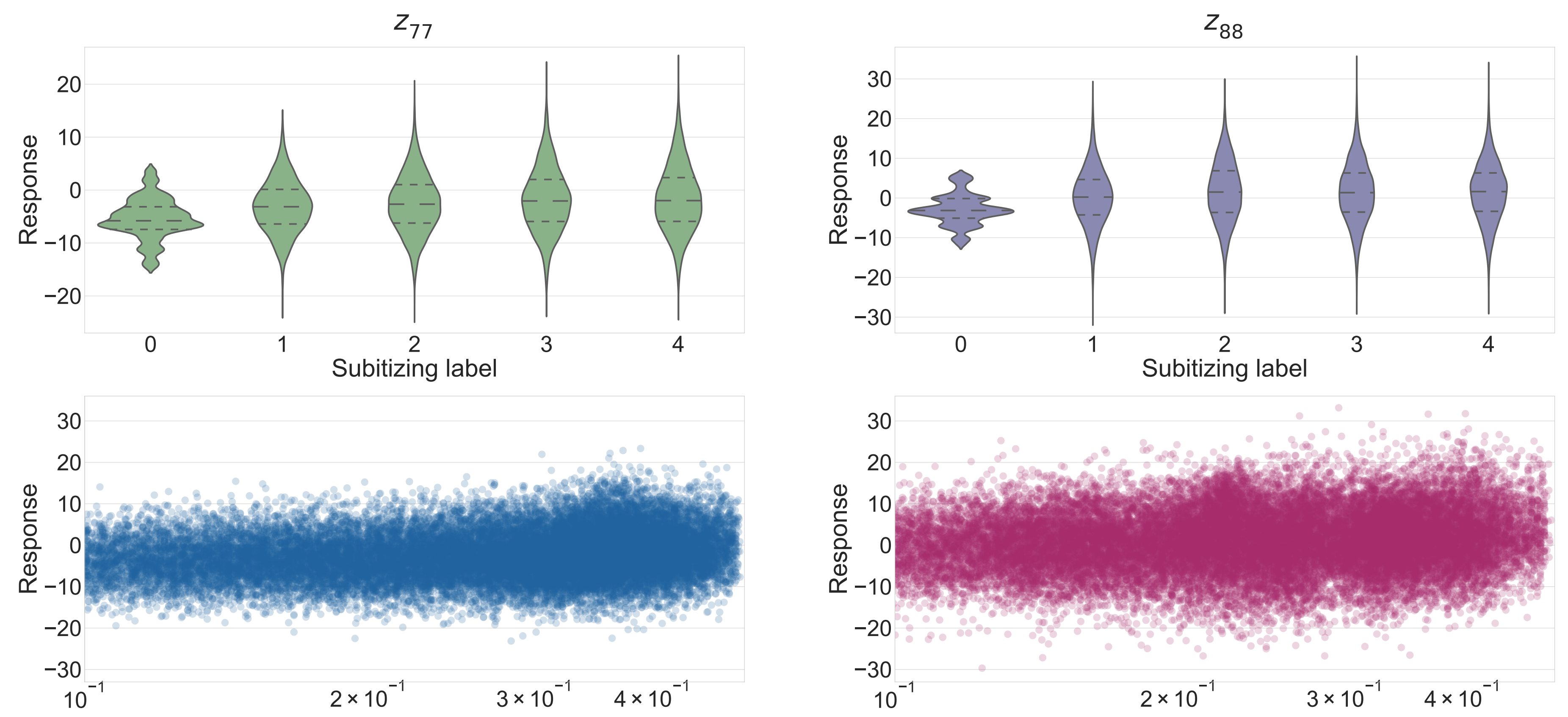
Augmentation of the SOS dataset with **synthetic data** was required to familiarize the model with an extensive distribution of object types and the spatial configurations thereof. Images are synthesized by cut-pasting objects onto natural image backgrounds with various random image transformations applied.

SUBITIZING TASK PERFORMANCE

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 fixing it's parameters, and subsequently training a simple **softmax classifier** that is latent representations of images with corresponding count labels. Performance is reported over the entire withheld SOS test set as count average precision (%).

SIZE-INVARIANT NUMEROSITY DETECTORS



In alignment with studies on biological neural networks, an analysis of the learned representations revealed that numerosity is represented **invariant to cumulative object area**. The results were obtained by performing a linear regression over a controlled synthetic dataset aimed at moderating object size and visual variation in object and background classes.

[1] Ivilin Stoianov and Marco Zorzi. Emergence of a 'visual number sense' in hierarchical generative models. *Nature Neuroscience*, 15(2):194, 2012.

[2] Jianming Zhang, Shugao Ma, Mehrnoosh Sameki, Stan Sclaroff, Margrit Betke, Zhe Lin, Xiaohui Shen, Brian Price, and Radomír Měch. Salient object subitizing. *IJCV*, 124(2):169–186, 2017.

[3] Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. In *WACV*, 2017.