Modeling Subitizing with Variational Autoencoders

Rijnder Wever,
rien334@gmail.com
10801944

June 28, 2018

1 Introduction

Although various machine learning approaches dealing with the numerical determination of the amount of objects in images already exist, some research can be impartial to broader cognitive debate, resulting in models that are somewhat unhelpful for progressing the understanding of numerical cognition. Any approach to computational modeling of numerical cognition that aims to maintain biological plausibility should adhere to neurological findings about the general characteristics of cognitive processes related to numerical tasks. Essential to understanding the cognitive processes behind number sense is their perceptual origin, for so called *visual numerosity* has been posed as the fundamental basis for developmentally later kinds of number sense, such as that required for arithmetical thinking and other more rigorous concepts of number found in mathematics (Lakoff and Núñez, 2000, chap. 2; Piazza and Izard, 2009). Visual numerosity is the perceptual capability of many organisms to perceive a group of items as having either a distinct or approximate cardinality. Some specific characteristics of visual numerosity can be derived from it's neural basis. Nieder (2016) and Harvey et al. (2013) present research where topologically organized neural populations exhibiting response profiles that remained largely invariant to all sensory modalities except quantity were discovered. Specifically, the found topological ordering was such that aside from populations responding to their own preferred numerosity, they also showed progressively diminishing activation to preferred numerosities of adjacent populations, in a somewhat bell-shaped fashion (Nieder, 2016). This correspondence between visual quantity and topological network structure leads them to conclude that neural populations can directly (i.e. without interposition of higher cognitive processes) encode specific visual numerosities. This coding property of neurons participating in visual numerosity was found in humans and other animals alike (see Nieder, 2016; Harvey et al., 2013).

Notwithstanding the success of previous biologically informed approaches to modeling numerical cognition with artificial neural networks (Stoianov and Zorzi, 2012), more work is to be done in applying such models to natural images. The main reason behind pursuing natural images is improving the biological plausibility over previous approaches relying on binary images containing only simple geometric shapes (for examples, see Stoianov and Zorzi, 2012; Wu et al., 2018), given that natural images are closer to everyday sensations than binary images. Furthermore, any dataset with uniform object categories does not capture how visual number sense in animals is abstract in regard to the perceived objects (Nieder, 2016), implying that a model should be able show that it performs equally well between objects of different visual complexities. Another way in which biological plausibility will be improved is by guiding some algorithmic decisions on the previously described discovery of the direct involvement of certain neural populations with numerosity perception. Moreover, the properties of these neurons' encoding scheme provide interesting evaluation data to the visual numerosity encoding scheme used by our final model. Some important and closely related characteristics of visual numerosity that constrain our approach, but hopefully improve the biological plausibility of the final model are:

1. Visual number sense is a purely *automatic* appreciation of the sensory world. It can be characterized as "sudden", or as visible at a glance (Dehaene, 2011, p. 57; Zhang et al., 2016a). *Convolutional neural networks* (CNNs) not only showcase excellent performance in extracting visual features from complicated natural images (Mnih et al., 2015; Krizhevsky et al., 2012; for visual number sense and CNNs see Zhang et al., 2016a), but are furthermore functionally inspired by the visual cortex of animals (specifically cats, see LeCun and Bengio, 1995). CNNs mimicking aspects of the animal visual cortex thus make them an excellent candidate for modeling automatic neural coding by means of numerosity percepts.

2. The directness of visual number entails that no interposition of external processes is required for numerosity perception, at least no other than lower-level sensory neural processes. More appropriately, the sudden character of visual number sense could be explained by it omitting higher cognitive processes, such as conscious representations (Dehaene, 2011, p. 58 indeed points to types of visual number sense being pre-attentive) or symbolic processing (visual numerosity percepts are understood non-verbally, Nieder, 2016). Furthermore, the existence of visual sense of number in human newborns (Lakoff and Núñez, 2000, chap. 1), animals (Davis and Pérusse, 1988) and cultures without exact counting systems (Dehaene, 2011, p. 261; Franka et al., 2008) further strengthens the idea that specific kinds of sense of number do not require much mediation and can purely function as an interplay between perceptual capabilities and neural encoding schemes, given the aforementioned groups lack of facilities for abstract reasoning about number (see Everett, 2005, p. 626; Lakoff and Núñez, 2000, chap. 3, for a discussion on how cultural facilities such as fixed symbols and linguistic practices can facilitate the existence of discrete number in humans). Indeed, Harvey et al. (2013) show that earlier mentioned neural populations did not dis-

play their characteristic response profile when confronted with Arabic numerals, that is, symbolic representations of number. These populations thus show the ability to function separately from higher-order representational facilities. Visual number sense being a immediate and purely perceptual process implies that our model should not apply external computational techniques often used in computer vision research on numerical determination task such as counting-by-detection (which requires both arithmetic and iterative attention to all group members, see Zhang et al., 2016a,b) or segmenting techniques (e.g. Chattopadhyay et al., 2016). Instead, we want to our model to operate in an autonomous and purely sensory fashion.

3. Relatedly, visual sense of number is an emergent property of hierarchically organized neurons embedded in generative learning models, either artificial or biological (Stoianov and Zorzi, 2012; the brain can be characterized as building a predictive modeler, or a "Bayesian machine", Knill and Pouget, 2004; Pezzulo and Cisek, 2016). The fact that visual number sense exist in animals and human newborns suggests that it is an implicitly learned skill learned at the neural level, for animals do not exhibit a lot of vertical learning, let alone human newborns having received much numerical training. Deemed as a generally unrealistic trope of artificial learning by AI critics (Dreyfus, 2007) and research into the human learning process (Zorzi et al., 2013a), modeling visual number necessitates non-researcher depended features. This will restrict the choice of algorithm to so called *unsupervised* learning algorithms, as such an algorithm will learn its own particular representation of the data distribution. Given their ability to infer the underlying stochastic representation of the data, i.e. perform in autonomous feature determination, *Variational Autoencoders* (VAEs) seem fit to tackle this problem (section 3.1 details their precise working). Moreover, VAEs are trained in an unsupervised manner similar to how, given appropriate circumstances, visual numerosity abilities are implicitly learned skills that emerge without "labeled data". Another interesting aspect of VAEs is their relatively interpretable and overseeable learned feature space, which might tell something about how it deals with visual numerosity, and thus allows us to evaluate the properties of the VAE's encoding against biological data.

Unfortunately, no dataset fit for visual numerosity estimation task similar to Stoianov and Zorzi (2012) satisfied above requirements (sizable collections of natural image with large and varied, precisely labeled objects groups are hard to construct), forcing present research towards *subitizing*, a type of visual number sense which had a catered dataset readily available. Subitizing is the ability of many animals to immediately perceive the number of items in a group without resorting to counting or enumeration, given that the number of items falls within the subitizing range of 1-4 (Kaufman et al., 1949; Davis and Pérusse, 1988) Most of the research above was conducted on approximate numerical cognition, but the aforementioned characteristics of visual sense of number hold equally well for a more distinct sense of number such as subitizing. Similarly, subitizing is suggested to be a parallel pre-attentive process in the visual system (Dehaene, 2011, p. 57), the visual system likely relying on it's ability to recognize holistic patterns for a final subitizing count (Jansen et al., 2014; Dehaene, 2011, p. 57; Piazza et al., 2002). This means that the "sudden" character of subitizing is caused by the visual system's ability to process simple geometric configurations of objects in parallel, whereby increasing the size of a group behind the subitizing range deprives perceiving this group of it's sudden and distinct numerical perceptual character for this would strain our parallelization capabilities too much. The difference in perceptual character is due to a recourse to enumeration techniques (and possibly others) whenever the subitizing parallelization threshold is exceeded, which differ from suddenness in being a consciously guided (i.e. attentive), patterned type of activity.

Present research therefore asks: how can artificial neural networks be applied to learning the emergent neural skill of subitizing in a manner comparable to their biological equivalents? To answer this, we will first highlight the details of our training procedure by describing a dataset constructed for modeling subitizing and how we implemented our VAE algorithm to learn a representation of this dataset. Next, as the subitizing task is essentially an image classification task, a methodology for evaluating the unsupervised VAE model's performance on the subitizing classification task is described. We demonstrate that the performance of our unsupervised approach is comparable with supervised approaches using handcrafted features, although performance is still behind state of the art supervised machine learning approaches due to problems inherent to the particular VAE implementation. Finally, measuring the final models robustness to changes in visual features shows the emergence of a property similar to biological neurons, that is to say, the VAE's encoding scheme supports numerosity percepts invariant to visual features other than quantity.

2 Related Work

Visual Number Sense. As previously described, Stoianov and Zorzi (2012) applied artificial neural networks to visual numerosity estimation, although without using natural images. They discovered neural populations concerned with numerosity estimation that shared multiple properties with biological populations participating in similar tasks, most prominently an encoding scheme that was invariant to the cumulative surface area of the objects present in the provided images. Present research hopes to discover a similar kind of invariance to surface area. Likewise, we will employ the same scale invariance test, although a successful application to natural images already shows a fairly abstract representation of number, as the objects therein already contain varied visual features.
Some simplicity of the dataset used by Stoianov and Zorzi (2012) is due their use of the relatively computationally expensive Restricted Boltzmann Machine (RBM) (with the exception of exploiting prior knowledge of regularities in the probability distribution over the observed data, equation (20.6) from Goodfellow et al., 2016, shows that computational cost in RBMs grows as a multiple of the size of it's hidden and

observed units). Given developments in generative algorithms and the availability of more computational power, we will therefore opt for a different algorithmic approach (see section 3.1) that will hopefully scale better to natural images.

Salient Object Subitizing Dataset. As seen in figure 1, the goal of the *Salient Object Subitizing* (SOS) dataset as defined by Zhang et al. (2016a) is to clearly show a number of salient objects that lies within the subitizing range. As other approaches often perform poor on images with complex backgrounds or with a large number of objects, Zhang et al. (2016a) also introduce images with no salient objects, as well as images where the number of salient objects lies outside of the subitizing range (labeled as "4+"). The dataset was constructed from an ensemble of other datasets to avoid potential dataset bias, and contains approximately 14K natural images (Zhang et al., 2016a).

Figure 1: Example images from the SOS dataset

3 Methods

3.1 Variational Autoencoder

VAEs are part of the family of autoencoder algorithms, owing this title to the majority of their structure consisting of an encoder and a decoder module (Doersch, 2016) (see figure 2 for the schematics of an autoencoder). In an regular autoencoder, the encoder module learns to map features from data samples $X \in \mathbb{R}^n$ into latent variables $z \in \mathbb{R}^m$ often so that $m \ll n$ and thus performs in dimensionality reduction, while the decoder function learns to reconstruct latent variables z into $X' \in \mathbb{R}^n$ such that X' matches X according to some predefined similarity measure (Liou et al., 2014). Reducing the input to be of much lower dimensionality forces the autoencoder to learn only the most emblematic regularities of the data, as these will minimize the reconstruction error. The latent space can thus be seen as an inferred hidden feature representation of the data.

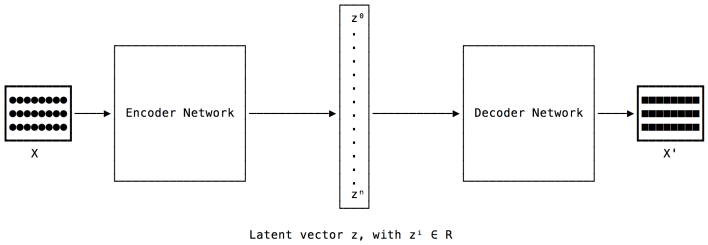


Figure 2: Schematic architecture of an autoencoder

Where VAEs primarily differ from regular autoencoders is that rather than directly coding data samples into some feature space, they learn the parameters of a distribution that represents that feature space. Therefore, VAEs perform stochastic inference of the underlying distribution of the input data, instead of only creating some efficient mapping to a lower dimensionality that simultaneously facilitates accurate reconstruction. Now provided with statistical knowledge of the characteristics of the input, VAEs can not only perform reconstruction, but also generate novel examples that are similar to the input data based on the inferred statistics. The ability to generate novel examples makes VAEs a *generative* algorithm.

The task of the encoder network in a VAE is to infer the mean and variance parameters of a probability distribution of the latent space $\mathcal{N}(z \mid \mu(X), \Sigma(X))$ such that samples z drawn from this distribution facilitate reconstruction of X (Doersch, 2016). Novel sampled z vectors can then be fed into the decoder network as usual. $\mu(X)$ and $\Sigma(X)$ are constrained to roughly follow a unit Gaussian by minimizing the Kullback-Leibler divergence (denoted as $\mathcal{KL}$) between $\mathcal{N}(0, I)$ and $\mathcal{N}(z \mid \mu(X), \Sigma(X))$, where $\mathcal{KL}$ measures the distance between probability distributions. Normally distributed latent variables capture the intuition behind generative algorithms that they should support sampling latent variables that produce reconstructions that are merely *similar* to the input, and not necessarily accurate copies (Doersch, 2016). Furthermore, optimizing an arbitrary distribution would be intractable, thus VAEs need to rely on the fact that given a set of normally distributed variables $S = \{s_1, ...s_\mathbf{n}\}$ with $S \in \mathbb{R}^\mathbf{n}$ and any sufficiently complicated function $f(s_i)$ (such as a neural network), there exists a mapping $f : S \mapsto S'$ from which we can generate any arbitrary distribution $P(X) \in \mathbb{R}^\mathbf{n}$ with $S' \sim P(X)$ (Doersch, 2016).

Therefore, the optimization objectives of a VAE become (also see figure 4 of Doersch, 2016):

1. $\mathcal{KL}[\mathcal{N}(\mu(X), \Sigma(X))||\mathcal{N}(0, I)]$
2. Some reconstruction loss. Within visual problems, plain VAEs can for example minimize the binary cross entropy (BCE) between X and X'.

Objective **(1)** grants VAEs the ability to generate new samples from the learned distribution, partly satisfying the constraint outlined in the introduction whereby visual numerosity related skills are shown to emerge in generative learning models. To fully satisfy this constraint, the final architecture uses deep neural networks for both the encoder and decoder module (see figure X for the VAE architecture), making the implementation an hierarchical model as well. As a VAEs latent space encodes the most important features of the data, it is hoped the samples drawn from the encoder provide information regarding it's subitizing performance (see section 4.1). For a complete overview of implementing a VAE, refer to Kingma and Welling (2013) and Doersch (2016).

3.2 Deep Feature Consistent Perceptual Loss

Because the frequently used pixel-by-pixel reconstruction loss measures in VAEs do not necessarily comply with human

perceptual similarity judgements, Hou et al. (2017) propose optimizing the reconstructions with help of the hidden layers of a pretrained deep CNN network, because these models are particularly better at capturing spatial correlation compared to pixel-by-pixel measurements (Hou et al., 2017). Additionally, CNNs haven proven to model visual characteristics of images deemed important by humans, by being able to for example perform complex image classification tasks (Krizhevsky et al., 2012). The ability of the proposed *Feature Perceptual Loss* (FPL) to retain spatial correlation should reduce the noted blurriness (**?**) of the VAE's reconstructions, which is especially problematic in subitizing tasks since blurring merges objects which in turn distorts subitizing labels. Hou et al. (2017) and present research employ VGG-19 (Simonyan and Zisserman, 2014) as the pretrained network Φ, trained on the large and varied ImageNet (Russakovsky et al., 2015) dataset. FPL requires predefining a set of layers $l_i \in L$ from pretrained network Φ, and works by minimizing the mean squared error (MSE) between the hidden representations of input x and VAE reconstruction $\bar{x}$ at every layer l_i. Aside from the $\mathcal{KL}$-divergence, the VAE's second optimization objective is now as follows:

$$\sum_{l \in L} = \text{MSE}(\Phi(x)^l, \Phi(\bar{x})^l)$$

The intuition behind FPL is that whatever some hidden layer l_i of the VGG-19 network encodes should be retained in the reconstruction $\bar{x}$, as the VGG-19 has proven to model important visual characteristics of a large variety of image types. In Hou et al. (2017)'s and our experiments $L = \{$`relu1_1, relu2_1, relu3_1`$\}$ resulted in the best reconstructions. One notable shortcoming of FPL is that although the layers from VGG-19 represent important visual information, it is a known fact that the first few layers of deep CNNs only encode simple features such as edges and lines (i.e they support contours), which are only combined into more complex features deeper into the network (Liu et al., 2017; FPL's authors Hou et al., 2017, note something similar). This means that the optimization objective is somewhat unambitious, in that it will never try to learn any other visual features (for examples, refer to Liu et al., 2017, Fig. 6.) aside from what the set of predefined layers L represents. Indeed, although contour reconstruction has greatly improved with FPL, the reconstruction of detail such as facial features shows less improvement. Although Hou et al. (2017) show a successful application of FPL, they might have been unaware of this shortcoming due to using a more uniform dataset consisting only of centered faces. For a comparison between FPL and BCE-based reconstruction measures on the SOS dataset, refer to figure 5.

3.3 Hybrid Dataset

We follow Zhang et al. (2016a) in pre-training our model with synthetic images and later fine-tuning on the SOS dataset. However, some small chances to their synthetic image pre-training setup are proposed. First, the synthetic dataset is extended with natural images from the SOS dataset such that the amount of examples per class per epoch is always equal (hopefully reducing problems encountered with class imbalance, see section 4.1.2). Another reason for constructing a hybrid dataset was the fact that the generation process of synthetic images was noted to produce 1. fairly unrealistic looking examples and 2. considerably less suitable than natural data for supporting subitizing performance (Zhang et al., 2016a). A further intuition behind this dataset is thus that the representation of the VAE must always be at least a little supportive of natural images, instead of settling on some optimum for synthetic images. A final reason for including natural images is that any tested growth in dataset size during pre-training resulted into lower losses. The ratio of natural to synthethic images is increased over time, defined by a bezier curve with parameters $u_0 = 0, u_1 = -0.01, u_2 = 0.02, u_3 = 1.0$ shown in figure 3. We grow the original data size by roughly 8 times, pre-training with a total of 80000 hybrid samples per epoch. Testing many different parameters for the hybrid dataset was not given much priority as the total loss seemed to shrink with dataset expansion and training and testing a full model was time expensive.

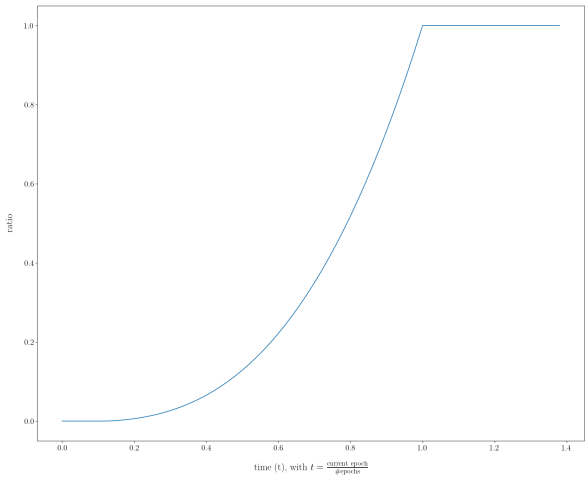


Figure 3: Bezier curve defining the ratio of natural images over syntethic images at time t

Synthetic images are generated by pasting cutout objects from THUS10000 (Cheng et al., 2015) onto the SUN background dataset (Xiao et al., 2010). The subitizing label is aquired by pasting an object N times, with $N \in [0, 4]$. For each paste, the object is transformed in equivalent manner to Zhang et al. (2016a). However, subitizing is noted be more difficult when objects are superimposed, forcing recource to external processes as counting by object enumeration (Dehaene, 2011, p. 57.), implying that significant paste object overlap should be avoided. Zhang et al. (2016a) avoid object overlap by defining a threshold $tin[0, 1]$ whereby an object's visible pixels $P_{visible}$ and total pixel amount P_{total} should satisfy $P_{visible} > t{*}P_{total}\}$. For reasons given above, we define the object overlap threshold as $t = \sum_{n=0}^{N} 0.5 + n_i * 0.1$ with $N \in [0, 4]$ compared to Zhang et al. (2016a)'s static $t = 0.5$, as VAEs are espcially prone produce blurry reconstructions (Hou et al., 2017; Larsen et al., 2015), which requires extra care with overlapping objects as

to not distort class labels. Refer to figure 4 for examples of generated synthetic images.

Figure 4: Synthetic images generated for the pre-training stage

4 Experiments

4.1 Hidden Representation Classifier

4.1.1 Classifier architecture

To asses whether the learned latent space of the VAE showcases the emergent ability to perform in subitizing, a two layer fully-connected net is fed with latent activation vectors z_{X_i} created by the encoder module of the VAE from an image X_i, and a corresponding subitizing class label Y_i, where X_i and Y_i are respectively an image and class label from the SOS training set. Both fully-connected layers contain 160 neurons. Each of the linear layers is followed by a batch normalization layer (Ioffe and Szegedy, 2015), a ReLU activation function and a dropout layer (Srivastava et al., 2014) with dropout probability x and y respectively. A fully-connected net was chosen because using another connectionist module for read-outs of the hidden representation heightens the biological plausibility of the final approach (Zorzi et al., 2013b). Additionaly, Zorzi et al. (2013b) note that the appended connectionist classifier module be concieved of as a cognitive response module supporting a particular behavioral task, although the main reason behind training this classifier is to asses it's performance against other algorithmic data.

4.1.2 Class imbalance

Class imbalance is a phenomenon encountered in datasets whereby the number of instances belonging to on or more classes is significantly higher than the amount of instances belonging to any of the other classes. Although there is no consensus on an exact definition of what constitutes a dataset with class imbalance, we follow Fernández et al. (2013) in that given over-represented class c_m the number of instances N_{c_i} of one the classes c_i should satisfy $N_{c_i} < 0.4 * N_{c_m}$ for a dataset to be considered imbalanced. For the SOS dataset, $N_{c_0} = 2596$, $N_{c_1} = 4854$, $N_{c_2} = 1604$ $N_{c_3} = 1058$ and $N_{c_4} = 853$, which implies that c_0 and c_1 are *majority classes*, while the others should be considered *minority classes*. Most literature makes a distinction between three general algorithm-agnostic approaches that tackle class imbalance (for a discussion, see Fernández et al., 2013). The first two rebalance the class distribution by altering the amount of examples per class. However, class imbalance can not only be conceived of in terms of quantitative difference, but also as qualitative difference, whereby the relative importance of some class is weighted higher than others (e.g. in classification relating to malignant tumors, misclassifying malignant examples as nonmalignant could be weighted stonger than other misclassifications) Qualitive difference might be relevant to the SOS dataset, because examples with overlapping (i.e. multiple) objects make subitizing inherently more difficult (see section 3.3), and previous results on subitizing show that some classes are more difficult to classify than others (Zhang et al., 2016a).

1. *Oversampling techniques* are a particularly well performing set of solutions to class imbalance. Oversampling alters the class distribution by producing more examples of the minority class, for example generating synthetic data that resembles minority examples (e.g. He et al., 2008; Chawla et al., 2002), resulting in a more balanced class distribution.
2. *Undersampling techniques.* Undersampling balances the class distribution by discarding examples from the majority class. Elimination of majority class instances can for example ensue by removing those instances that are highly similar (e.g. Tomek, 1976)
3. *Cost sensitive techniques.* Cost sensitive learning does not alter the distribution of class instances, but penalizes misclassification of certain classes. Cost sensitive techqniques are especially useful for dealing with minority classes that are inherently more difficult (or "costly") to correctly classify, as optimisation towards easier classes could minimize cost even in quantitatively balanced datasets if the easier classes for example require lesser representational resources of the learning model.

An ensemble of techniques was used to tackle the class imbalance in the SOS dataset. First, slight random under-sampling with replacement of the two majority classes (c_0 and c_1) is performed (see Lemaître et al., 2017), reducing their size by ~10%. Furthermore, as in practice many common sophisticated under- and oversampling techniques (e.g. data augmentation or outlier removal, for an overview see Fernández et al. (2013)) proved largely non-effective, a cost-sensitive class weighting was applied. The uneffectiveness of quantive sampling techniques is likely to be caused by that in addition to the quantitative difference in class examples, there is also a slight difficulty factor whereby assesing the class of latent vector z is significantly if belongs to c_2 or c_3 versus any other class, for these two classes require rather precise contours to discern invidual objects, even more so with overlapping objects, while precise contours remain hard for VAEs given their tendency to produce blurred reconstructions (Larsen et al., 2015). The classifier network therefore seems inclined to put all of its representational power towards the easier classes, as this will result in a lower total cost, whereby this inclination will become even stronger as the quantitative class imbalance grows. The class weights for cost sensitive learning are set according to the quantitative class imbalance ratio (similair to section 3.2 in Fernández et al., 2013), but better accuracy was obtained by slightly altering the relative difference between the weights by raising them to some power n. In our experiments, $n = 3$ resulted in a bal-

ance between high per class accuray scores and aforementioned scores roughly following the same shape as in other algorithms, which hopefully implies that the classifier is able to generalize in a manner comparable to previous approaches. For the SOS dataset with random majority class undersampling, if $n \gg 3$ the classifier accuracy for the majority classes shrinks towards chance, and, interestingly, accuracy for the minority classes becomes comparable to the state of the art machine learning techqniques.

5 Results & Discussion

5.1 Variational Autoencoder Performance

After pre-training for 102 epochs, and fine-tuning on the SOS dataset for 39 epochs, we found that the loss of our VAE did not shrink anymore. For the specific purpose of subitizing, we can see that using FPL loss is beneficial (indeed, that is what we found when comparing the two models in the classification task described in section 4.1) The reconstructions of a plain VAE and a VAE that uses FPL as it's reconstruction optimization objective are show in figure 5. To get an idea of what sort of properties the latent space of the VAE encodes, refer to figure 6.

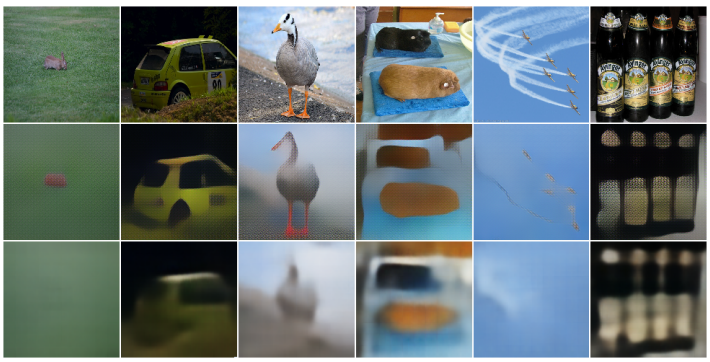


Figure 5: Comparison between BCE and FPL reconstruction loss measures. The top row are the original images from the SOS dataset, and the other two images are reconstructions made by using FPL and BCE loss, respectively. (the VAE using BCE was pre-trained for 40 epochs, and fine-tuned for 35 epochs)

Although the reconstructions show increased quality of contour reconstruction, there are a few reoccurring visual disparities between original and reconstruction. First of, novel patterns often emerge in the reconstructions, possibly caused by a implementational glitch, or a considerable difference in tested datasets (FPL is frequently paired with the CelebA (Liu et al., 2015) dataset). Datasets other than the SOS dataset showed slightly better performance, indicating that the SOS dataset is either to small, too varied or requires non standard tweaking for FPL to work in it's current form. Most of the improvement in more uniform datasets came from the fact that the VAE learned to create more local patterns to give the appearance of uniformly colored regions, but upon closer inspection placed pixels of colors in a grid such that they gave the appearance of just one color, similar to how for example LED screens function. Another reconstructional problem is that small regions such as details are sometimes left out, which could possibly distort class labels (object might start to resemble each other less if they lose detail).

5.2 Subitizing Read-Out

Accuracy of the `zclassifier` (i.e. the classifier as described in section 4.1.1 concerned with classification of latent activation patterns to subitizing labels) is reported over the withheld SOS test set. We report best performances using a slightly different VAE architecture than the one described in section 3.1 (scoring a mean accuracy of 40.4). The main difference between the VAE used in this experiment and the one that is used throughout the rest of this research is it places intermediate fully connected layers (with size 3096) between the latent representation and the convolutional stacks. Accuracy scores of other algorithms were copied over from Zhang et al. (2016a). For their implementation, refer to Zhang et al. (2016a).

	0	1	2	3	4+	mean
Chance	27.5	46.5	18.6	11.7	9.7	22.8
SalPry	46.1	65.4	32.6	15.0	10.7	34.0
GIST	67.4	65.0	32.3	17.5	24.7	41.4
zclasifier	76	49	40	27	30	44.4
SIFT+IVF	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

The subitizing performance of the VAE is comparable to highest scoring non-machine learning algorithm, and performs worse overall than the CNNs trained by Zhang et al. (2016a). This can be explained by a number of factors. First of all, the `CNN_ft` algorithm used by Zhang et al. (2016a) has been pretrained on a large, well tested and more varied dataset, namely ImageNet (Russakovsky et al., 2015), which contains $\approx$ 1300x more images. Additionally, their model is capable of more complex representations due its depth and the amount of modules it contains (the applied model from Szegedy et al., 2015, uses 22 compared to the 12 in our approach). Moreover, all their algorithms are trained in a supervised manner, providing optimization algorithms such as stochastic gradient descent with a more directly guided optimization objective, an advantage over present research's unsupervised training setup.

5.3 Qualitative Analysis

Artificial and biological neural populations concerned with visual numerosity support quantitative judgements invariant to object size and, conversely, some populations detect object size without responding to quantity, indicating a separate encoding scheme for both properties (Stoianov and Zorzi, 2012; Harvey et al., 2013). Analogously, we tested whether our VAE's latent representation contained dimensions z_i encoding either one of these properties. To test this, we first created a dataset with synthetic examples containing N objects ($N \in [0, 4]$, with N uniformly distributed over the dataset) and

Figure 6: Reconstructions of the image in the top-left made by slightly increasing the response value of the VAE's latent representation z, at different individual dimensions z_i. Some dimensions give you a slight idea of what types of information they encode (e.g. a light source at a location)

corresponding cumulative area values A that those N objects occupied (measured in pixels, with A normally distributed over the dataset[1]). The object overlap threshold was set to 1 for each example, to reduce noise induced by possible weak encoding capabilities and reasons outlined in section 3.3. As visualisations showed that each dimension z_i encodes more than one type of visual feature (see figure 6) special care was undertaken to reduce z_i's response variance by only generating data with 15 randomly sampled objects from the object cut-out set, and one random background class from the background dataset (performance is reported on the "sand deserts" class, which contains particularly visually uniform examples). A dimension z_i is said to be able to perform as either a numerical or area detector when regressing it's response over novel synthetic dataset ($n = 33280$) supports the following relationship between normalized variables A and N (Stoianov and Zorzi, 2012):

$$z_i = \beta_1 \log(N) + \beta_2 \log(A) + \varepsilon \quad (\text{with } N \in [0, 4]) \qquad (1)$$

The regression was accomplished with linear regression algorithms taken from Newville et al. (2016) (Levenberg–Marquardt proved best). The criteria set by Stoianov and Zorzi (2012) for being a good fit of (1) are **1)** the regression explaining at least 10% of the variance ($R^2 \geq 0.1$) **2)** and a "ideal" detector of some property should have a low ($\mid \beta_i \mid < 0.1$) regression coefficient for the complementary property. We slightly altered criteria **1** to fit our training setup. The complexity of the SOS dataset in comparison to the binary images used by Stoianov and Zorzi (2012) requires our model to encode a higher variety of information, meaning that any fit is going to have more noise as no dimension z_i has one role (see figure 6 for an overview). Moreover, the synthetic data we use for the regression includes more complex information than the dataset used by Stoianov and Zorzi (2012). Nevertheless, we still found a small number of reoccurring detectors of A and N, with $R > 0.065 \pm 0.020$ (all z_i with $R < 0.033$ resulted in an average $R = 0.06 \pm 0.01$). Due to randomisation in the fitting process (synthetic examples are randomly generated at each run) the role distribution varied slightly with each properties being encoded by about 1-2 dimensions, out of the total of 182 (anymore would indicate an unlikely redundancy, given that the small latent space should provide an efficient encoding scheme). Some latent dimensions that provide a better fit of (1) exist, but don't satisfy criteria **2**. An interesting note is that whenever the regression showed multiple dimensions encoding area, they either exhibited positive or negative responses (i.e. positive or negative regression coefficients) to area increase, in accordance with how visual numerosity might rely on a size normalisation signal, according to some theories on the neurocomputational basis for numerosity (see Stoianov and Zorzi, 2012, for a discussion). A large negative response (in contrast to a positive) to cumulative area might for example be combined with other response in the VAE's decoder network as an indicatory or inhibitory signal that the area density does not come from just one object, but from multiple.

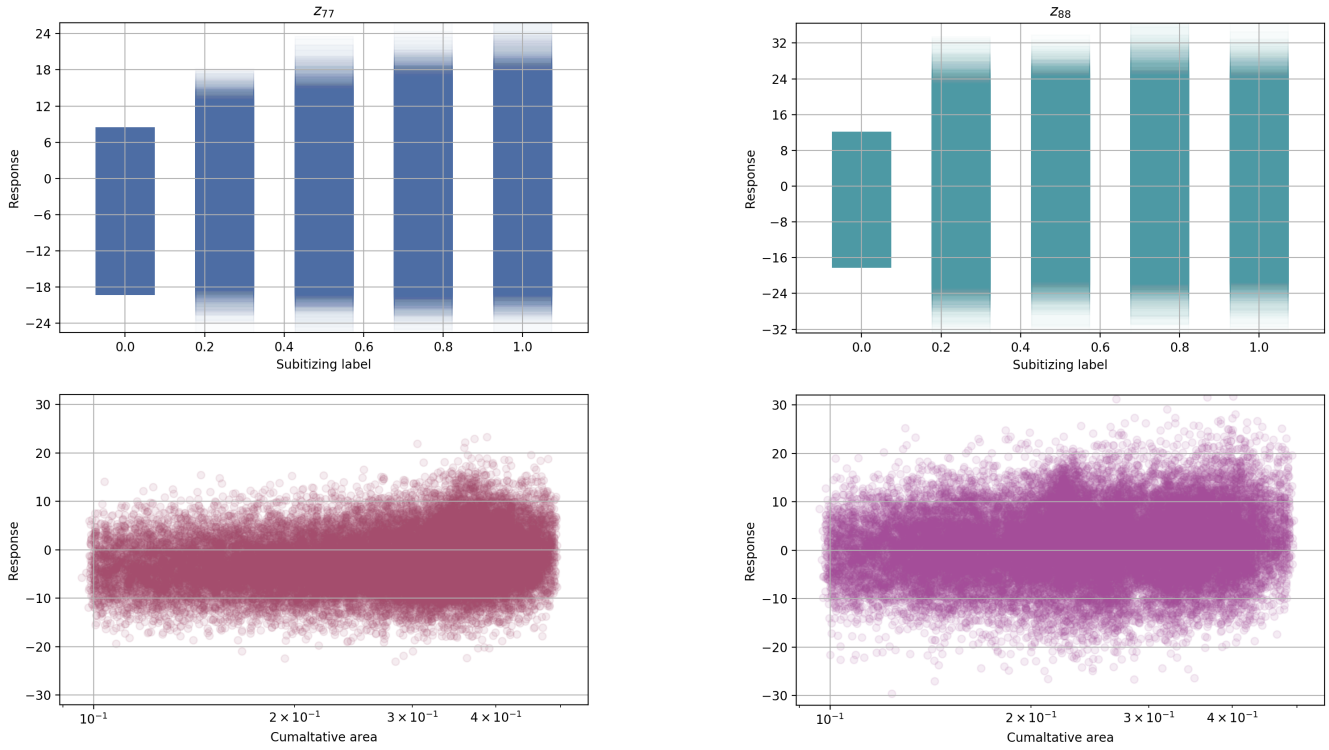
Figure 7 provides characteristic response profiles for dimensions encoding either cumulative area or a subitizing count. For the area dimension (figure 7b), extreme cumulative area samples bend the mean distribution either upwards or downwards, while the response distribution to cumulative area for numerosity encoding dimensions stays relatively centered. The cumalative area detector z_{88} also shows an increasing response value relative to an increase in cumaltive area, espcially in comparison to z_{77}. For numerosity dimension z_{77}, figure 7a shows that both the total response and the center of the response distribution increased with numerosity (note

[1] A uniform distribution of cumulative area might have worked better, but required algorithmic changes to the synthetic data generation process that were inhibited by the amount of time still available.

the transparent areas). In contrast, the dimension that was sensitive cumalative area shows a fairly static response to changes in subitizing count. With some extra time, the visual clarity and overall performance of this qualitative analysis could probably be greatly improved, given that only a short focus on reducing response variance increased R by almost a factor of 10 in some cases.

6 Conclusion

We described a setup for training a VAE on a subitizing task, while satisfying some important biological constraints. A possible consequence thereof is that our final model showcases properties also found in biological neural networks (as well as other artificial algorithms). Firstly, an ability to subitize emerged as an implicitly learned skill. Second of all, the learned encoding scheme indicates support for encoding numerosities without resorting to counting schemes relying to cumulative (objective) area, and conversely encodes cumulative area without using numerosity information, in accordance with previous (other?) comparable artificial models (Stoianov and Zorzi, 2012). However, more research is needed to asses , as more properties of input need to be varied (such as visual variation and the distribution of variables), there is room for improvement in the VAE's reconstructional abilities, i.e. efficiency of coding scheme. These two problems also indicate room for improvement in the subitizing classification task, which has the additional improvement of solving the class imbalance problem. Nevertheless, visual numerosity-like skills have emerged during the training of the VAE, showing the overall ability to perceive numerosity within the subitizing range without using information provided by visual features other than quantity. We can thus speak of a fairly abstract sense of number, as the qualitative analysis of the encoding yielded promising results over a large variation of images, whereby especially abstraction in regard to scale has been demonstrated.



(a) (b)

Figure 7: (a) shows a typical response profile for a numerosity detector ($R = 0.055$). Subitizing label N was normalized. (b) shows a typical response profile of dimension that encodes cumulative area while being invariance to numerosity information ($R = 0.056$). Cumulative area (A) was normalized and is displayed across a logarithmic scale. For visual convenience, examples with $A = 0$ were shifted next to lowest value of A in the dataset.

References

George Lakoff and Rafael E Núñez. Where mathematics comes from: How the embodied mind brings mathematics into being. *AMC*, 10:12, 2000.

Manuela Piazza and Véronique Izard. How humans count: numerosity and the parietal cortex. *The neuroscientist*, 15 (3):261–273, 2009.

Andreas Nieder. The neuronal code for number. *Nature Reviews Neuroscience*, 17(6):366–382, may 2016. doi: 10.1038/nrn.2016.40. URL https://doi.org/10.1038%2Fnrn.2016.40.

Ben M Harvey, Barrie P Klein, Natalia Petridou, and Serge O Dumoulin. Topographic representation of numerosity in the human parietal cortex. *Science*, 341(6150):1123–1126, 2013.

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

Xiaolin Wu, Xi Zhang, and Jun Du. Two is harder to recognize than tom: the challenge of visual numerosity for deep learning. *arXiv preprint arXiv:1802.05160*, 2018.

Stanislas Dehaene. *The number sense: How the mind creates mathematics*. OUP USA, 2011.

Jianming Zhang, Shuga Ma, Mehrnoosh Sameki, Stan Sclaroff, Margrit Betke, Zhe Lin, Xiaohui Shen, Brian Price, and Radomír Měch. Salient object subitizing. *arXiv preprint arXiv:1607.07525*, 2016a.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

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.

Yann LeCun and et al. Bengio, Yoshua. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Hank Davis and Rachelle Pérusse. Numerical competence in animals: Definitional issues, current evidence, and a new research agenda. *Behavioral and Brain Sciences*, 11(4): 561–579, 1988.

Michael C Franka, Daniel L Everettb, Evelina Fedorenkoa, and Edward Gibsona. Number as a cognitive technology: Evidence from pirahã language and cognitionq. *Cognition*, 108:819–824, 2008.

Daniel L. Everett. Cultural constraints on grammar and cognition in pirahã. *Current Anthropology*, 46(4):621–646, aug 2005. doi: 10.1086/431525. URL https://doi.org/10.1086%2F431525.

Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Unconstrained salient object detection via proposal subset optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5733–5742, 2016b.

Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R Selvaraju, Dhruv Batra, and Devi Parikh. Counting everyday objects in everyday scenes. *arXiv preprint arXiv:1604.03505*, 2016.

David C Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.

Giovanni Pezzulo and Paul Cisek. Navigating the affordance landscape: feedback control as a process model of behavior and cognition. *Trends in cognitive sciences*, 20(6):414–424, 2016.

Hubert L Dreyfus. Why heideggerian ai failed and how fixing it would require making it more heideggerian. *Philosophical psychology*, 20(2):247–268, 2007.

Marco Zorzi, Alberto Testolin, and Ivilin P. Stoianov. Modeling language and cognition with deep unsupervised learning: a tutorial overview. *Frontiers in Psychology*, 4, 2013a. doi: 10.3389/fpsyg.2013.00515. URL https://doi.org/10.3389%2Ffpsyg.2013.00515.

E. L. Kaufman, M. W. Lord, T. W. Reese, and J. Volkmann. The discrimination of visual number. *The American Journal of Psychology*, 62(4):498, oct 1949. doi: 10.2307/1418556. URL https://doi.org/10.2307%2F1418556.

Brenda RJ Jansen, Abe D Hofman, Marthe Straatemeier, Bianca MCW Bers, Maartje EJ Raijmakers, and Han LJ Maas. The role of pattern recognition in children's exact enumeration of small numbers. *British Journal of Developmental Psychology*, 32(2):178–194, 2014.

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.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

Cheng-Yuan Liou, Wei-Chen Cheng, Jiun-Wei Liou, and Daw-Ran Liou. Autoencoder for words. *Neurocomputing*, 139: 84–96, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu. Deep feature consistent variational autoencoder. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 1133–1141. IEEE, 2017.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Mengchen Liu, Jiaxin Shi, Zhen Li, Chongxuan Li, Jun Zhu, and Shixia Liu. Towards better analysis of deep convolutional neural networks. *IEEE transactions on visualization and computer graphics*, 23(1):91–100, 2017.

Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Marco Zorzi, Alberto Testolin, and Ivilin Peev Stoianov. Modeling language and cognition with deep unsupervised learning: a tutorial overview. *Frontiers in psychology*, 4:515, 2013b.

Alberto Fernández, Victoria López, Mikel Galar, María José del Jesus, and Francisco Herrera. Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-Based Systems*, 42:97–110, apr 2013. doi: 10.1016/j.knosys.2013.01.018. URL https://doi.org/10.1016%2Fj.knosys.2013.01.018.

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pages 1322–1328. IEEE, 2008.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Ivan Tomek. Two modifications of cnn. *IEEE Trans. Systems, Man and Cybernetics*, 6:769–772, 1976.

Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017. URL http://jmlr.org/papers/v18/16-365.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. Going deeper with convolutions. Cvpr, 2015.

Matthew Newville, Till Stensitzki, Daniel B Allen, Michal Rawlik, Antonino Ingargiola, and Andrew Nelson. Lmfit: non-linear least-square minimization and curve-fitting for python. *Astrophysics Source Code Library*, 2016.