UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Computer Science Curriculum

Tetiana Rabiichuk

Numerosity Sense in Artificial Neural Networks

Master's Thesis (30 ECTS)

Supervisor: Raul Vicente Zafra, PhD

Tartu 2022

Numerosity Sense in Artificial Neural Networks

Abstract:
The ability to approximately assess the number of objects in the set is observed in humans as well as animals. The mechanism of emergence of this ability is still an open research question. In this work, we consider approaching this question with the help of artificial neural networks from two perspectives: as an emergent property of interaction with the objects in the world through actions (as proposed by [KP20]), and as an emergent property of bottom-up projections of visual system (as proposed by [KJB$^+$21]). The first approach leads to topological organization of the embedding space of the network in a linear monotonic way with respect to cardinality of embedded samples that resembles "mental number line". The second approach leads to the detection of numerosity-sensitive artificial units with tuning properties that resemble tuning properties of real neurons recorded in monkey prefrontal cortex. Through a series of control experiments we demonstrate that representation that emerges in artificial units of both models does not disentangle abstract property of numerosity of a set from visual properties of objects constituting this set that are confounded with numerosity.

Keywords:
Number Sense, Deep Learning, Convolutional Neural Networks

CERCS:
P176 - Artificial Intelligence, S260 - Psychology, B640 - Neurology, neuropsychology, neurophysiology

Numbritaju tehisnärvivõrkudes

Lühikokkuvõte:
Võimet ligikaudselt hinnata hulgas olevate objektide arvu täheldatakse nii inimestel kui ka loomadel. Selle võime tekkimise mehhanism on endiselt lahtine uurimisküsimus. Käesolevas töös vaatleme sellele küsimusele lähenemist tehisnärvivõrkude abil kahest vaatenurgast: kui interaktsiooni maailma objektidega tegevuste kaudu esilekerkivat omadust (nagu pakkus välja [KP20]) ja kui visuaalse süsteemi alt-üles projektsioonide esilekerkivat omadust (nagu pakkus välja [KJB$^+$21]). Esimene lähenemine viib võrgu sisestusruumi topoloogilise korralduseni sisestavate näidiste kardinaalsuse suhtes lineaarsel monotoonsel viisil, mis meenutab "vaimset arvujoont". Teine lähenemine viib numbrite suhtes tundlike tehisüksuste tuvastamiseni, mille häälestusomadused sarnanevad ahvi prefrontaalses ajukoores registreeritud tõeliste neuronite häälestusomadustega. Kontrollkatsete seeria abil demonstreerime, et mõlema mudeli tehisüksustes ilmnev esitus ei lahuta hulga abstraktset arvulisuse omadust selle hulga moodustavate objektide visuaalsetest omadustest, mis on põimitud arvulisusega.

Võtmesõnad:
numbritaju, sügavõpe, konvolutsioonilised närvivõrgud

CERCS:
P176 - Tehisintellekt, S260 - Psühholoogia, B640 - Neuroloogia, neuropsühholoogia, neurofüsioloogia

Contents

1 Introduction

> *It must have required many ages to discover that a brace of pheasants and a couple of days were both instances of the number 2: the degree of abstraction involved is far from easy.*

Bertrand Russell, *Introduction to Mathematical Philosophy*, 1919

Number sense, which is an ability to effortlessly (without restoring to counting) approximately assess the number of the objects in the set (*cardinality*) is an ability that is attributed to humans (as early as early as 6 month age [XSG05]), as well as monkeys [NM07], birds [DN15], cats [TMRP70], and other animals [But22]. This ability is also considered to be a foundation for mathematical ability. However, whether this ability is developed or innate, and in the former case, what contributes to its emergence is still an open research question. Apart from psychophysics experiments in humans ([BAA17]) and electrophysiological studies in animals, scientists have attempted to approach the study of the numerosity perception with the help of mathematical modelling and artificial neural networks ([VF04], [DC93]). With the rapid advances in the field of deep learning, there has been a renewal in interest in approaching the question of numerosity sense with artificial neural networks. In this work, we are going to focus on two recent works.
The first work [KP20] focuses on the question: "what can facilitate the emergence of numerosity sense in a child in an unsupervised way". The authors propose a model of how the numerosity perception can be acquired by children through interaction with the world. The idea is based on the interaction between action and perception: the representation learned by the visual system is optimized to predict future actions. As a result of training to predict the action performed by a motor system of the child, the model learns a very peculiar embedding space: samples with different numerosities are embedded on a "number line", which is akin to the "number line" [ID08] representation attributed to humans. The authors claim that the network learned the concept of numerosity during training, and that the embedding space organizes samples with respect to the numerosity.
In the second work [KJB$^+$21] the authors report detecting populations of numerosity-sensitive artificial neurons, that exhibit properties resembling numerosity-sensitive neurons recorded in the monkey prefrontal cortex [NM07]). However, unlike previous works ([SZ12], [NVN19]), where the numerosity-selective neurons were shown to emerge as a result of a supervised training of the network, in [KJB$^+$21], the authors report detecting numerosity-sensitive neurons in randomly initialized untrained neural network.
Numerosity is an abstract property of the set, it requires the ability to abstract it from visual properties of the objects observed, hence it is interesting to look into its manifistation in artificial neural networks. In this thesis, we are performing additional experiments to

test the main claims of the aforementioned papers. We replicate the model from [KP20] and test the robustness of the learned representation in the embedding space with respect to the variability of the training data. We explore how persistent is the linear structure ('number line') that emerges in the embedding space and investigate the impact of the embedding space dimensionality on the structure of the learned representation. We also explored the impact of the input to the classification network on the structure of the embedding space. In order to test the claims of the [KJB$^+$21], we replicate the detection of numerosity-selective neurons and inspect the tuning propeties of the detected neurons. We add further control for the variables confound with the numerosity to the detection method, and observe how robust are the detected neurons to additional controls.

2 Background

2.1 Number sense and its characteristics

The ability to approximately assess the number of objects in the visual field is referred to as numerosity perception. It is important to note that numerosity perception is a different concept from counting. *When we count objects, we sequentially pay attention to all of the objects in the scene. We need to memorize the current object count and increase it as we encounter unobserved object. This is a conscious process, while numerosity estimation is a perception task, it does not require a conscious mental effort.*
Numerosity estimation in both humans and animals has several notable characteristics. Humans can very precisely and quickly tell the cardinality if the number of objects is below 4 (this range is called *subitizing range*), after that the precision of the estimation drops as the number of objects increases. When estimating which set of objects contains more objects, the estimation is more precise when both sets consist of smaller number of objects, while the precision drops when both sets have larger number of objects (Fechner-Weber law). Second, it is easier to discriminate between two sets if the difference between cardinalities is larger (distance effect). It is still an open research question how the numerosity perception ability is acquired.

2.2 Number neurons

In [NM07], the authors recorded from monkey brain while monkeys performed delayed match-to-numerosity task. Monkeys have been presented with a stimulus: a display of dots of a particular numerosity. After some delay they have been presented with two new stimuli, from which they had to select the one that contains the same number of dots as initial stimulus. Since the authors wanted their subjects to make the choice based on numerosity of the samples, rather then other properties of the stimulus, confound with numerosity (e.g. total area and density), they introduced control datasets for those correlations, where the value of this confound variable is fixed for all presented numerosities. The authors wanted to find neurons that would be selective to numerosity property of the set: their activity would be modulated by numerosity of the stimulus, rather then properies confound with numerosity. Hence, they run a two-way ANOVA on recorded data using numerosity $(1, 2, 4, 6, \ldots, 28, 30)$ and stimulus dataset (Standard, where no control for confound variable is introduced, and two control datasets: Same Total Area and Same Density) as factors. Detected *number-selective neurons* (see Figure 1) exhibited signature tuning properties: their average activity peaks at their **preferred numerosity** (numerosity that on average elicited the strongest response of this neuron), while it gradually decays as the distance between presented and preferred numerosity increases.

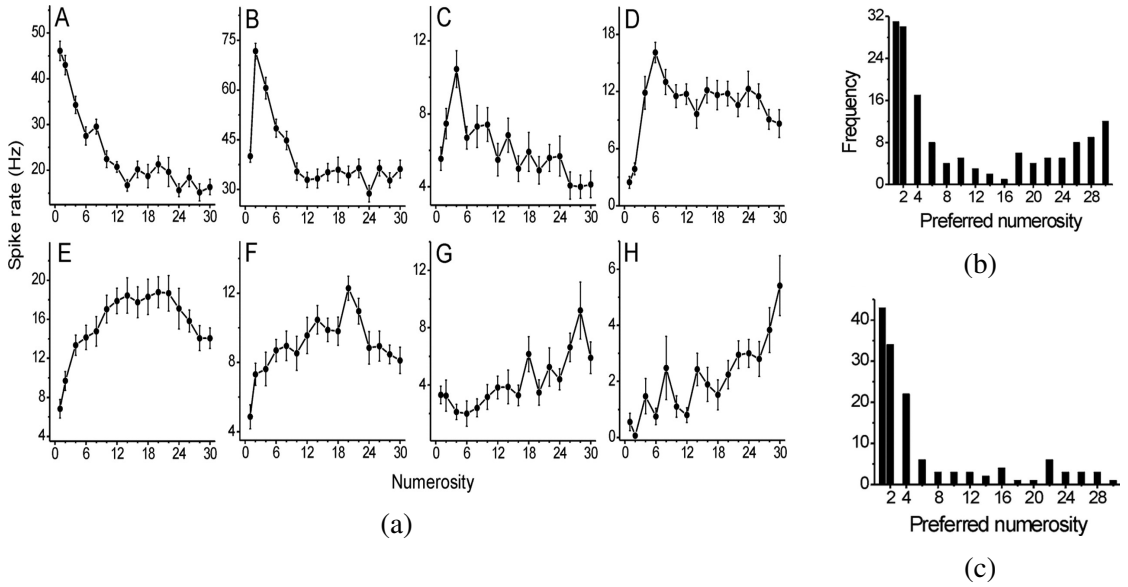
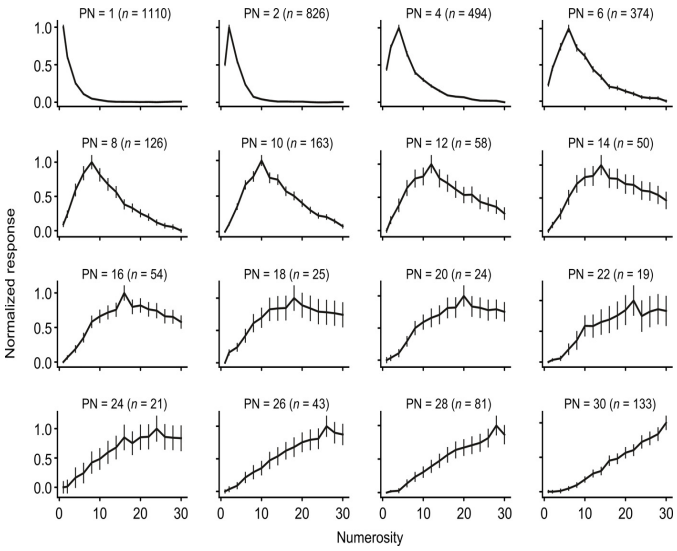


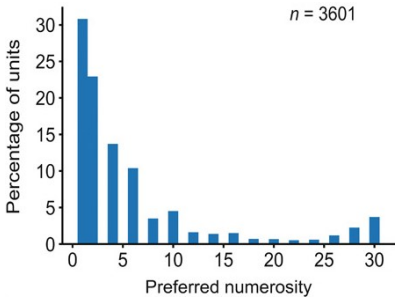
Figure 1. Numerosity-selective neurons in monkey prefrontal cortex. (a) Example tuning functions of sample numerosity-selective neurons from monkey prefrontal cortex with different preferred numerosities. (b) and (c) Distribution of preferred numerosity across detected numerosity-selective neurons during (a) sample period and (b) delay period.

2.2.1 Number neurons in artificial neural networks

[NM07] has inspired a series of work on searching for numerosity-selective neurons in artificial neural networks. In a recent work [NVN19] the authors study the activation of units of a hierarchical convolutional neural network (HCNN) [LB95] trained to classify instances of ImageNet [DDS$^+$09] dataset, which contains 1000 classes of natural image objects. After pre-training the network on ImageNet, the authors exposed the network to dataset of dots of different numerosities, akin to those used in [NM07]: one Standard Dataset, and two control datasets: Same Total Area Dataset and Same Convex Hull Dataset (aforementioned datasets are used in our work as well, please refer to the Datasets subsection for detailed description), and analysed the activations of artificial units in the final convolutional layer preceding fully-connected classification part of the model. The authors used the definition of numerosity-selective unit provided in [NM07]. Detected numerosity-selective artificial units exhibit tuning properties and distribution of preferred numerosity similar to the one obtained for real recorded neurons in [NM07] (see Figure 2).

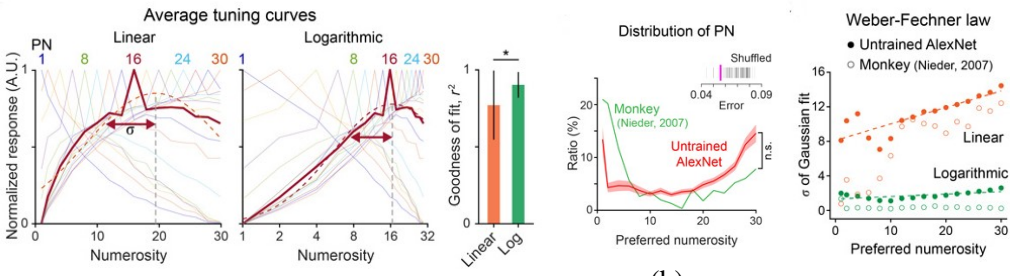


(a) Tuning curves of numerosity-selective network units.



(b) Distribution of preferred numerosities of the numerosity-selective network units.

Figure 2. Tuning properties of numerosity-selective neurons detected in final convolutional layer of HCNN pre-trained on ImageNet [NVN19].



(a) (b) (c)

Figure 3. Tuning properties of numerosity-selective neurons detected in final convolutional layer (Conv5) of untrained randomly initialized AlexNet architecture [KJB$^+$21]. (a) Tuning curves of numerosity-selective network units plotted on a linear (left) and logarithmic (right) scale. (b) Distribution of preferred numerosities of the numerosity-selective network units. (c) σ of a Gaussian fit to average tuning curve of units with different preferred numerosities.

Previously, artificial units with response profile akin to the one of numerosity-selective neurons found in the brain were reported as an emergent phenomena, e.g. after training a generative model on a dataset of dots of different numerosities ([SZ12]), or after training a HCNN on an unrelated task of classification of natural images ([LB95]). However, one may be curious if training is indeed needed for "number units" to emerge. This question was addressed in a recent work [KJB$^+$21], where the authors analysed activations of the last convolutional layer (Conv5) of an untrained randomly initialized AlexNet [KSH17] (see Figure 7 for architecture description) architecture using the methodology from [NM07] and [NVN19] outlined above. Surprisingly, even "naïve" untrained network contains units tuned to specific numerosity, and its tuning properties closely resemble tuning properties of real "number neurons" reported in [NM07] (see Figure 3).

3 Methods

3.1 "Manipulating brain" model

The model proposed in [KP20] aims to model the process of acquiring a numerosity sense in a child without explicit supervision (e.g. a parent explicitly telling the number of objects a child observes) through interaction with an objects in the world. Imagine a child playing with toys. A child observes the objects it interacts with, and it can either add a new toy (*put*), remove a toy (*take*) or change the positions of the objects in the scene (*shake*). The numerosity of the set of objects the child observes changes after put and shake operations, and is *invariant* with respect to shake operation. The authors model how the features extracted by the visual system of a child are optimized to predict the action that has just been performed (see Figure 4). The source code of the [KP20] is not currently publicly available, hence in our work we replicate this model using the description in the original paper. The representations learned by replicated model matches closely the ones reported in the paper. Our version of "manipulative brain model" is implemented in Python using the PyTorch [PGM+19] framework. As stimulus sets we utilized datasets with openly available source code, that were used in other works on numerosity perception ([NVN19], [BTZ21]) or their modifications. Those datasets are equivalent in complexity, or more diverse than the ones used in original paper.

3.1.1 Architecture

The model consists of two parts: encoder (plays the role of the perception) and a classifier that acts on top of the extracted features to predict the action performed by the motor system. Since the model needs to compare the state of the world *before* and *after* the action has been performed, the encoder is applied in parallel to corresponding inputs, hence the resulting architecture is a Siamese network [BBB+93]. The visual system is modeled by a modification of the AlexNet architecture [KSH17], that has been shown to learn representations close to the ones from higher level area of visual system: Inferior temporal (IT) cortex of humans and monkeys [KRK14]. The architecture of the encoder is summarized in Figure 4. The first two convolutional layers are initialized with weights of the AlexNet pre-trained on ImageNet [DDS+09]. Note, that since the weights of the first 3 layers of the encoder network are initialized from the weights of the pre-trained AlexNet, the input to the network has 3 color channels to be compatable with required input, however, since samples in stimulus sets used in this work are greyscale, all 3 channels contain the same information.

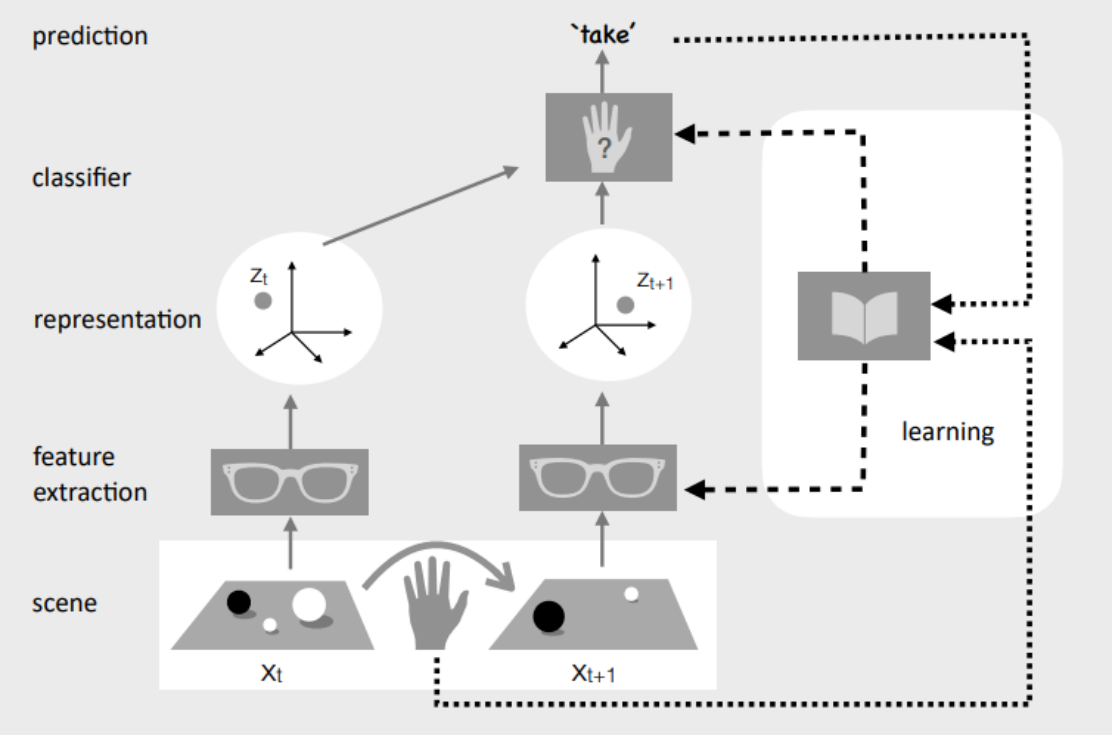


Figure 4. Schematic representation of the 'Manipulative brain" model [KP20]

3.1.2 Dataset generation procedure

The model is trained on a sequence of pairs of images, where each next image is obtained by randomly selecting one of the available actions for the current image: when the current numerosity is minimal considered numerosity, only *put* or *shake* actions are allowed, similarly, for the maximum allowed numerosity only *take* or *shake* actions are available. Same as in the original paper, in our implementation *put* and *take* operations are actually superposition of *put* and *shake*, and *take* and *shake* operations respectively. See Figure 6 for a sample training sequence.

3.1.3 Training procedure

The the objective function minimized during training was negative loglikelihood (NLL). The model has been trained using Adam optimizer [KB14] with learning rate $1e-4$. For all of the experiments, the training setup was kept the same as in the original paper: the model has been trained for 30 epochs, with 30 mini batches per epoch, where each mini batch consists of 180 image pairs. Hence, $30 \times 30 \times 180 = 162000$ image pairs were observed by the model during training.

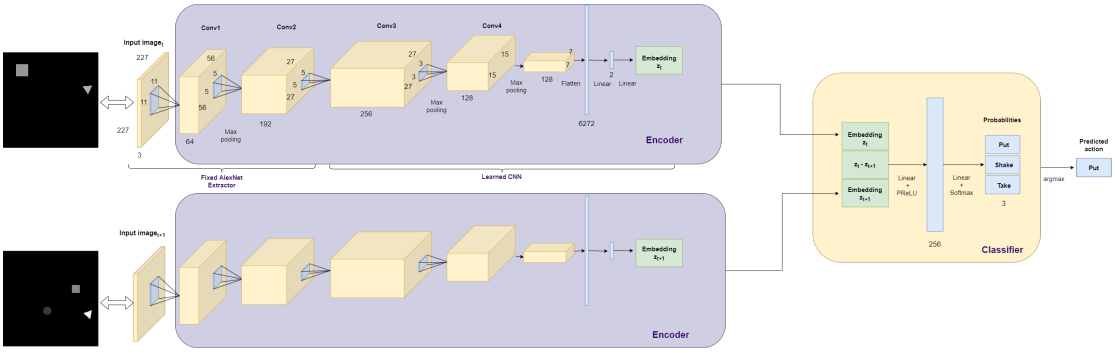


Figure 5. Schematic representation of the Manipulative Brain model architecture.

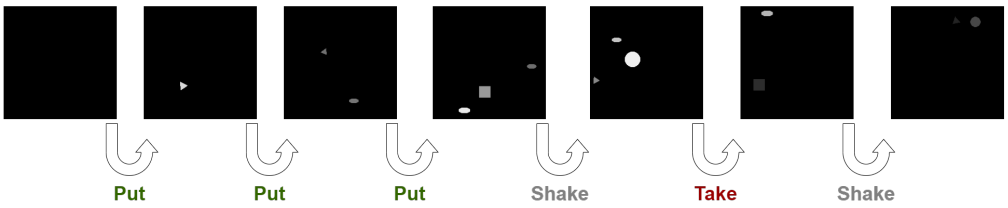


Figure 6. Example of training sequence generated from Variable Size, Shape and Contrast dataset for manipulating brain model.

3.2 Numerosity-sensitive neurons in randomly initialized network

3.2.1 Architecture

Following [KJB+21], we have used AlexNet [KSH17] architecture for our analysis (see Figure 7). In our analysis we have used a version of AlexNet architecture available from torchvision.models subpackage of PyTorch framework. Note that the default initialization method in Pytorch is He initialization [HZRS15], which is a default random initialization in deep learning toolbox of MATLAB (MathWorks Inc.) which was used in the original paper.

3.2.2 Detection of numerosity-selective units

Following previous works ([KJB+21], [NVN19]), two-way ANOVA analysis with *numerosities* $1, 2, 3, \ldots 30$ and *stimulus set* (Standard Dataset, Same Total Area, Same Convex Hull) as factors was applied to analyse the activation of units of layers of interest. In addition to analyzing responses of units in Conv-5 layer, in or work we have also analyzed the responses of preceding convolutional layers Conv-2, Conv-3 and Conv-4 after ReLU activation function of AlexNet architecture. Artificial unit is considered

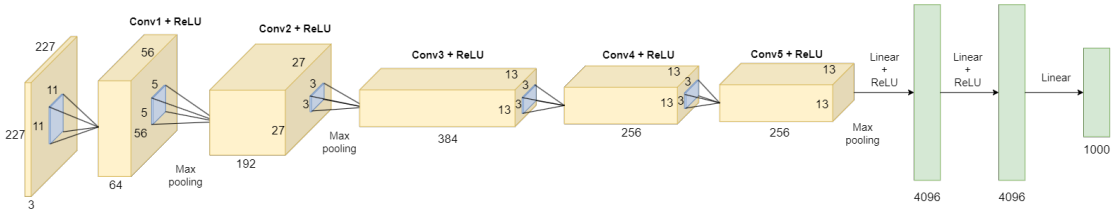


Figure 7. Schematic representation of the AlexNet architecture, units of which were analysed for numerosity selectivity.

numerosity-selective if its response shows a significant effect for numerosity, and no significant effect for stimulus set or interaction of numerosity and stimulus set ($p < 0.01$).

3.2.3 Tuning curves of numerosity-selective neurons

For each detected numerosity-selective neuron the numerosity that elicited its maximum average response was considered its **preferred numerosity** (PN). For each numerosity-selective unit its tuning curve was computed by averaging its activity over all presentations of the given numerosity ($150 \times \#$ of datasets). In order to compute average tuning curves of units tuned to each numerosity, tuning curves of units with the same preferred numerosity were averaged and the resulting average response was mapped to $[0, 1]$ range via min-max normalization. We make use of ANOVA implementation from statsmodels [SP10] Python module.

3.3 Datasets

In this subsection we describe the stimulus sets that have been used in our work.In all of the datasets for all numerosities the objects are guaranteed to be fully located within the boundaries of the image, and not intersect with other objects present in the image.
Uniform Dots dataset ([BTZ21]) consists of images of 224×224 pixels in size containing dots of the same radius and maximum intensity. To generate samples of this dataset, we have utilized source code of this dataset available on GitHub [BC21b]. See Figure 8 for examples of generated samples.

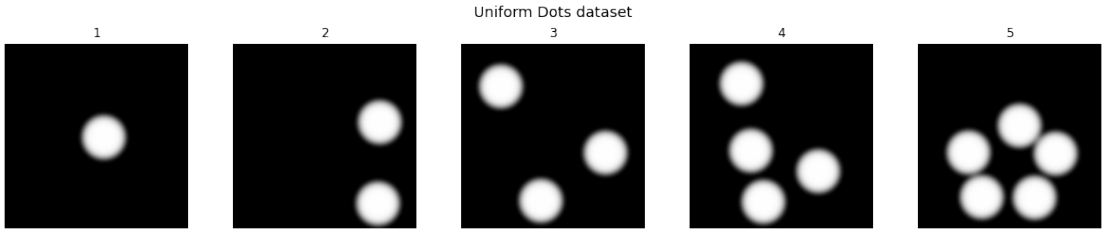


Figure 8. Examples of generated samples for Uniform Dots datset.

Non-Uniform Dots (variable size) ([BTZ21]) Is a modification of Uniform Dots dataset, where samples contain dots of variable size and maximum contrast, such that expected value of cumulative area of samples of different numerosities is constant. We have utilized source code available on GitHub [BC21a]. See Figure 9 for sample examples.

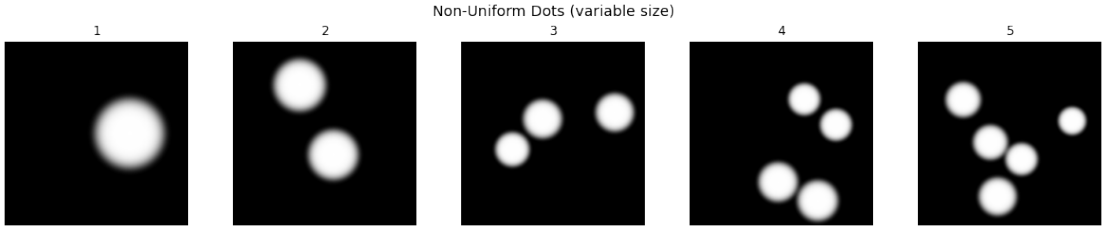


Figure 9. Examples of generated samples for Non-Uniform Dots (variable size).

Non-Uniform Dots (variable size and contrast) Is a modification of Non-Uniform Dots (variable size) dataset where global contrast of the image was randomly varied in range $1 - 100\%$. Note, that in this dataset all objects still have the same intensity. See Figure 10 for sample examples.

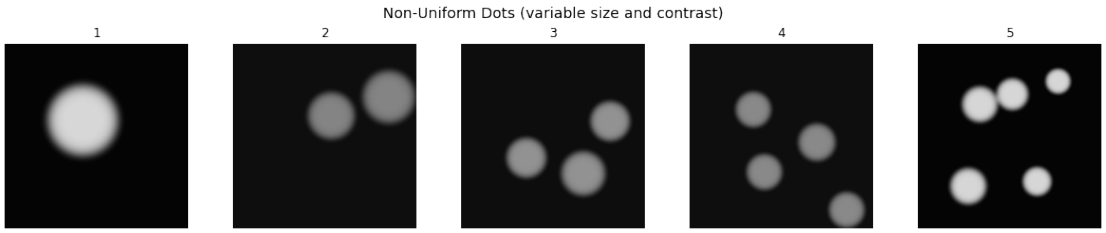


Figure 10. Examples of generated samples for Non-Uniform Dots (variable size and contrast).

Standard Dataset ([NVN19], [KJB$^+$21]) consists of images of size 227×227 that consists of circles of nearly identical radius: mean = 7, std = 0.7. To generate samples of this dataset we utilized code provided in [KJB$^+$21] source code. See Figure 11 for sample examples.

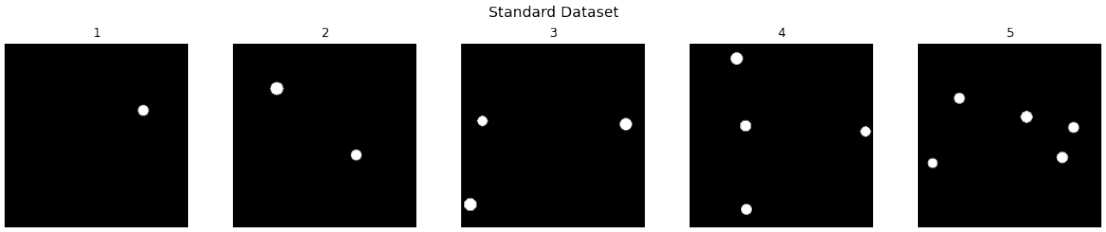


Figure 11. Examples of generated samples for Standard Dataset.

Same Total Area ([NVN19], [KJB$^+$21]) Is a control dataset, where the total area for samples of different numerosities is fixed at 1200 px, and average distance between dots kept in range 90-100 pixels. Since the total area is kept fixed, the area of individual objects decreases as the numerosity increases. See Figure 12.

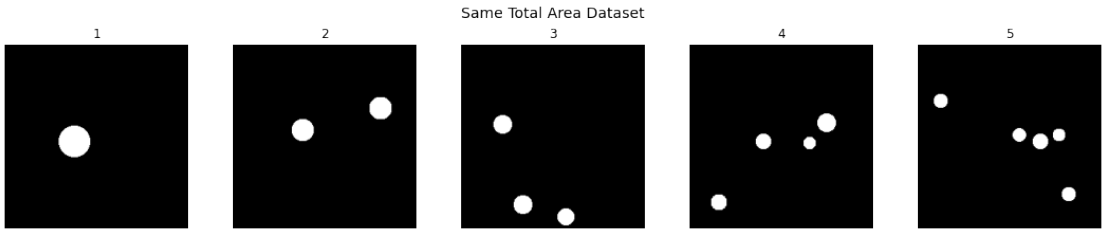


Figure 12. Examples of generated samples for Same Total Area Dataset.

Same Convex Hull ([NVN19], [KJB$^+$21]) Is a control dataset where objects were

located in the convex hull with the shape of the regular pentagon with circumference of 647 pixels (for numerosities ≥ 4), the shape of each dot was selected randomly from the shapes of: a circle, a rectangle, an ellipse and a triangle with equal probability. See Figure 13.

Figure 13. Examples of generated samples for Same Convex Hull Dataset.

Same Total Perimeter We have modified source code of Same Total Area dataset, to generated samples of numerosities $1, 2, 3, \ldots, 30$, such that all samples have total contour length ≈ 550. See Figure 14.

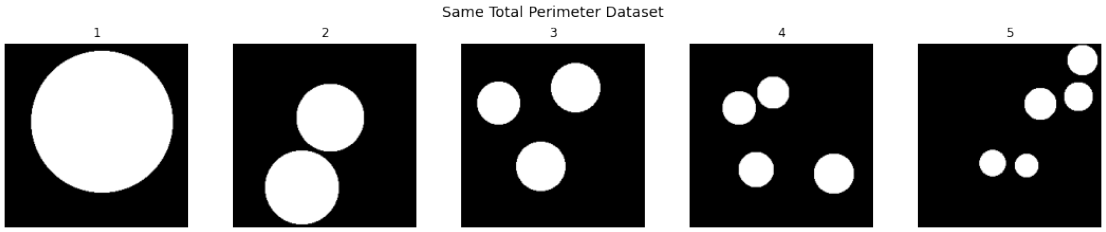


Figure 14. Examples of generated samples for Same Total Perimeter Dataset.

Variable Size and Contrast In a modification of Standard Dataset, such that the radius of individual circles varied in range $7 - 14$, and the intensity of individual objects varied in range $10 - 100\%$. See Figure 15.

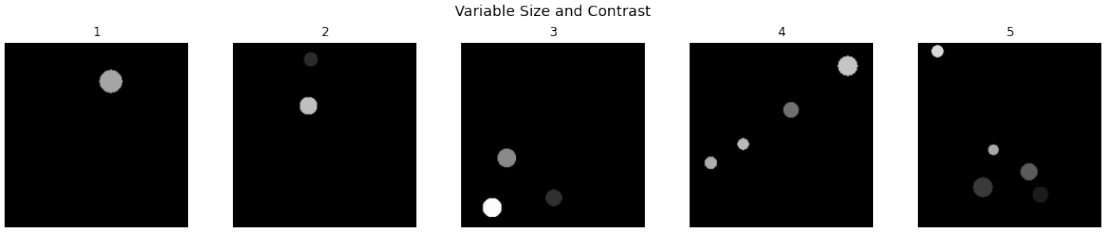


Figure 15. Examples of generated samples for Variable Size and Contrast Dataset.

Variable Size, Shape and Contrast is a modification of Same Convex Hull dataset, where the intensity of individual objects varied in range $10 - 100\%$, and object size was determined by the radius of the circumscribed circle, that varied in range $7 - 14$ pixels. See Figure 16

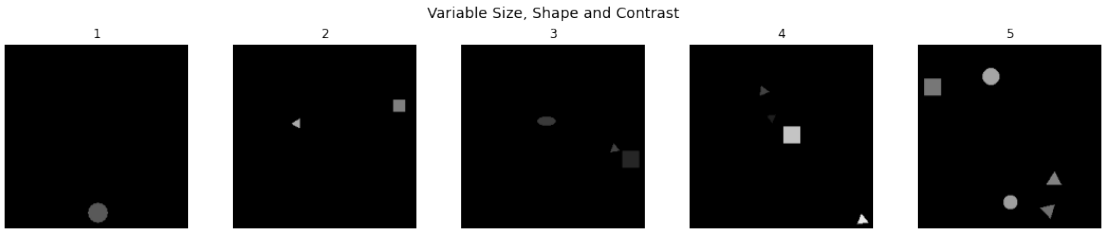


Figure 16. Examples of generated samples for Variable Size, Shape and Contrast Dataset.

Morphing Ellipse We have gradually morphed a circle with radius $r = 9$ into an ellipse, by generating an ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} < 1$ with parameters $a = r,\ b = \alpha r$ or $a = \alpha r,\ b = r$, where parameter α - morphing coefficient characterizes elongation of ellipse (see Figure 17). We have considered morphing coefficients in range $(0, 5)$ with 0.2 step. The location and orientation (vertical or horizontal) of generated ellipse was randomized. We have generated 150 samples for each morphing coefficient.

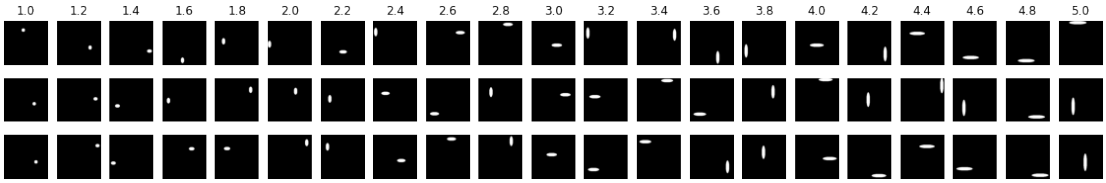


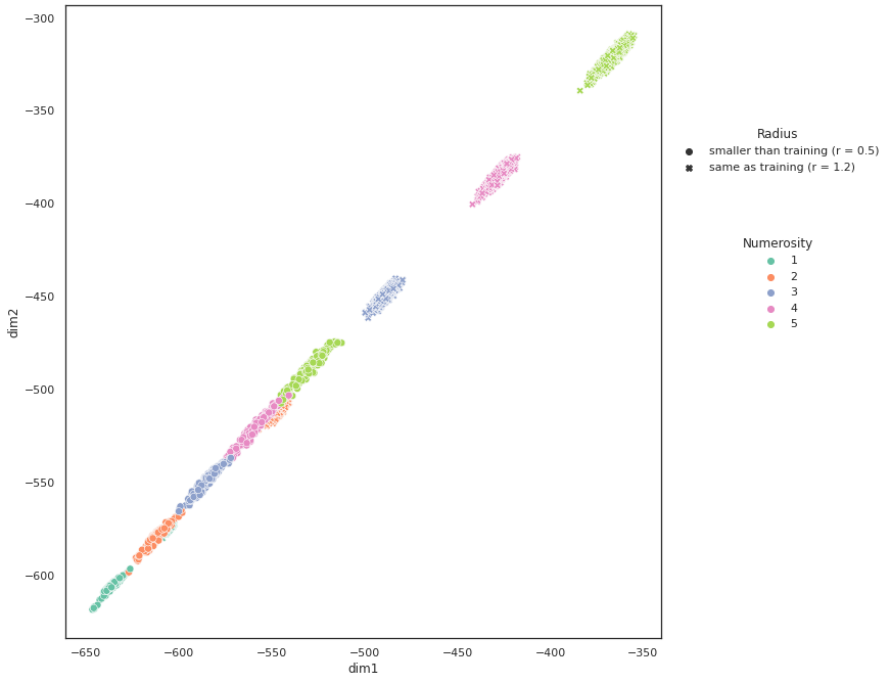
Figure 17. Examples of generated samples for Morphing Ellipse dataset per each morphing coefficient α.

4 Results

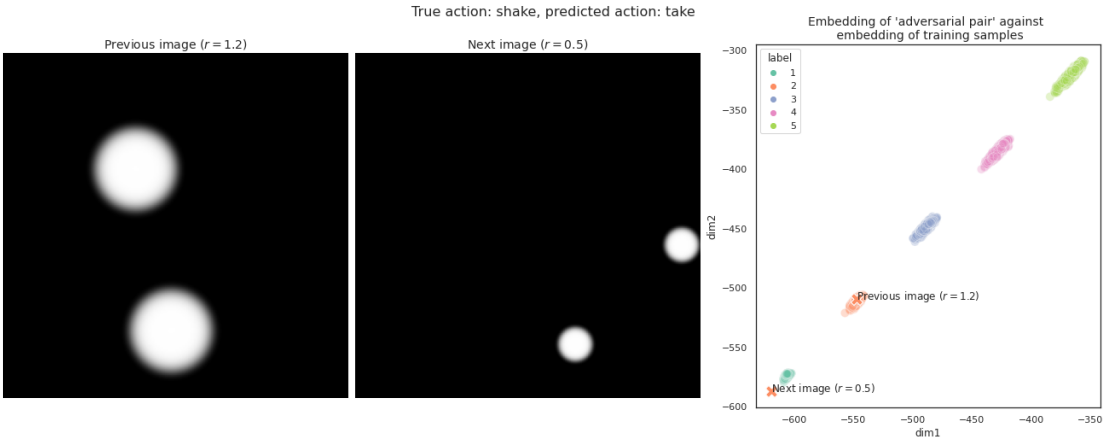
4.1 Results for manipulative brain model

4.1.1 Robustness of the learned 'number line' representation depends on the variability of the dataset

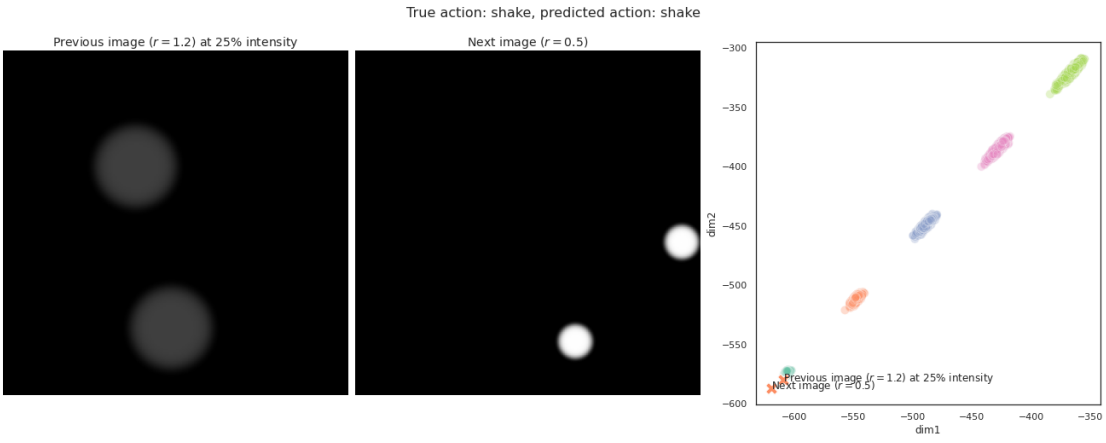
First, we investigated how the variability present in the dataset impacted the *number line* representation that emerged in the embedding space of the model. We observed that in case when the numerosity clearly correlates (e.g. Uniform Dots dataset) with other confound variables (e.g. total area, total intensity, etc.), the model did not learn to abstract the concept of numerosity from its correlates as demonstrated in Figure 18. If the model organized embedding space with respect to numerosity, we would expect samples of different radius but the same numerosity to be mapped close together. However, as one can observe in Figure 18a, samples with smaller radius, not observed during training, are shifted in the direction of the smaller numerosity with respect to the embedding of training samples. Figures 18b and 18c show how the embedding of samples impacts the prediction. In Figure 18b we observe that the samples of smaller radius is mapped close to cluster of samples of numerosity 1, and the model outputs the wrong action prediction *take*. Figure 18c demonstrates that the embedding is influenced by the total intensity of a sample. If we sufficiently reduce the intensity of the sample with larger radius, the model embeds it sufficiently close to the sample of smaller radius, and, as a result, the model outputs the correct action prediction *shake*, though both samples are mapped to the cluster of the wrong numerosity. If one adds the variabilty in radius size into the training set, the model correctly maps samples of the same numerosity but different radius close together in the embedding space (see Figure 19a), but it is still susceptible to intensity manipulation (see Figure 19b). Adding more degrees of variability to the training set makes the embedding more robust with respect to 'naive' adversarial examples as the ones mentioned before, as this is one of the ways to tell the model which properties of the objects are irrelevant for the numerosity estimation.



(a)

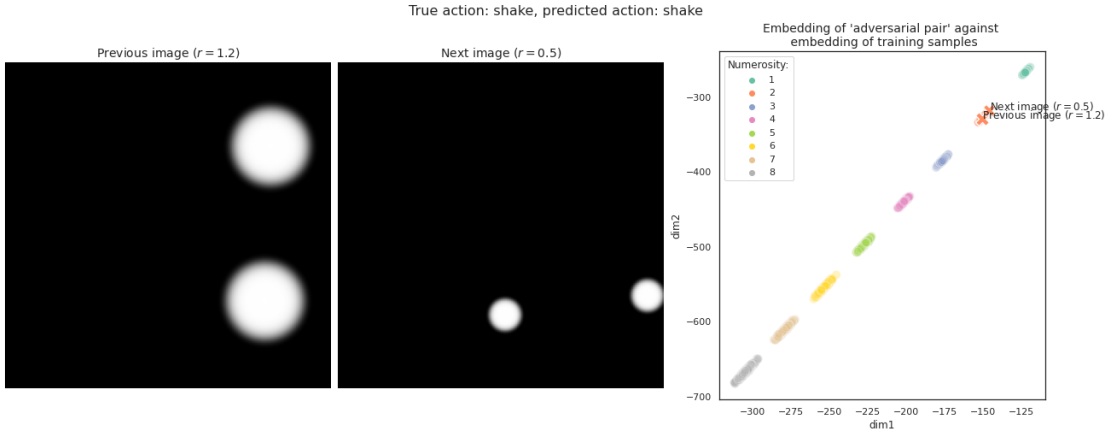


(b)

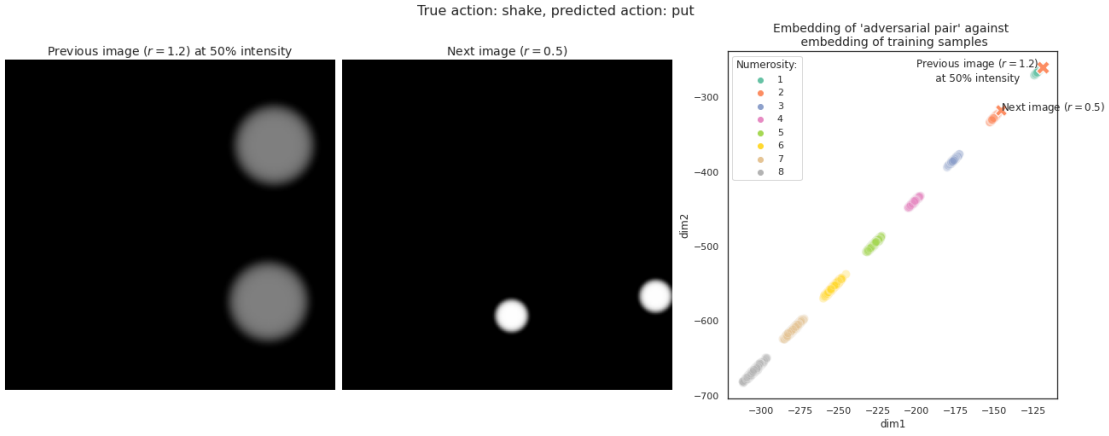


(c)

Figure 18. Embedding visualization of the model trained on Uniform Dots dataset.

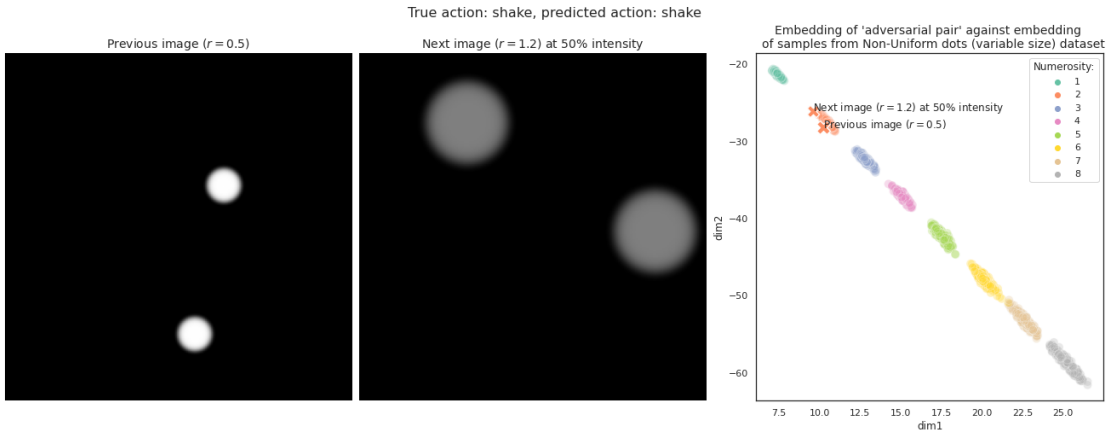


(a)

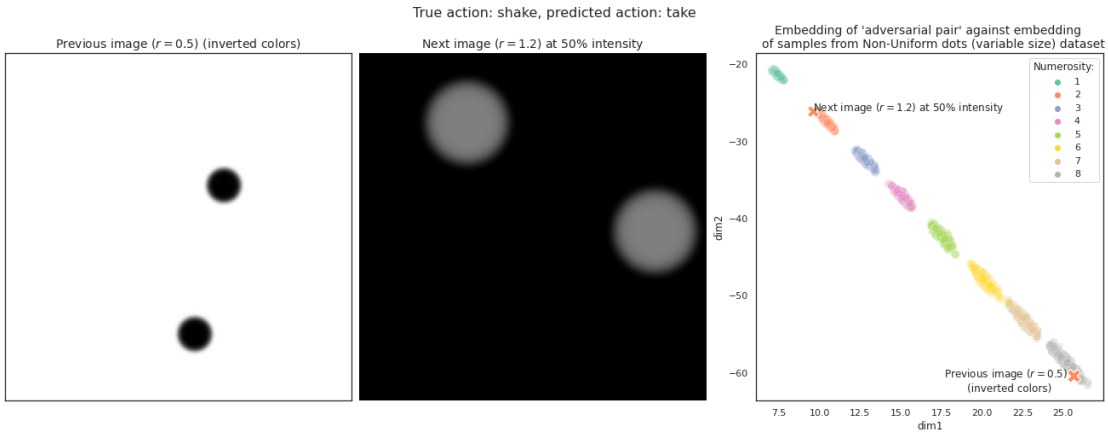


(b)

Figure 19. Embedding visualization of the model trained on Non-Uniform Dots (variable size).



(a)



(b)

Figure 20. Embedding visualization of the model trained on Non-Uniform Dots (variable size and contrast).

Please refer to the Appendix for additional experiments on robustness of the embedding of the model with respect to confound variables.

4.1.2 'Number line' representation persists in higher dimensions

We have varied the dimensionality of the embedding space of the model and visualized its projection using linear dimensionality reduction method: Principal Component Analysis (PCA), and a non-linear dimensionality reduction technique: t-Distributed Stochastic Neighbor Embedding (t-SNE) (see Figure 21). We have observed that the monotonical

organization of the clusters of different numerosities persists as the dimensionality of the embedding increases. Interestingly, in most of the cases, PCA projection of numerosities up to 3 were located at a different angle to the embedding of larger numerosities. Another salient feature of embedding projection was that t-SNE projection for all considered dimensions was intrinsically one-dimensional.

4.1.3 Effect of the structure of the input to classifier on the topology of learned embedding

We have also investigated the effect of the structure of the input to the classifier on the resulting embedding topology. Recall, that the classifier obtains its input in form of $(z_t, z_{t+1}, z_t - z_{t+1})$, where z_t and z_{t+1} are the embeddings of the first and second images of the pair. We have observed that the monotonic and linear structure of the embedding space is preserved if the difference $z_t - z_{t+1}$ is not provided as an input to classifier, as well as if this is the only input to the classifier (see Figure 22).

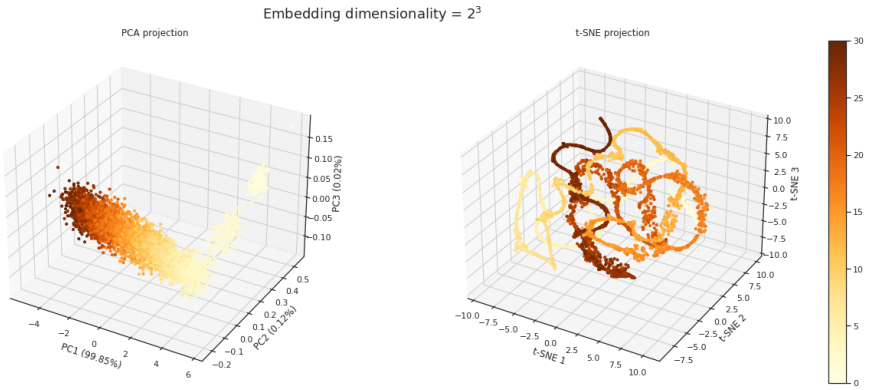
4.1.4 Network is able to extrapolate the 'number line' representation in both directions from the training range

In the original paper, the authors demonstrate that the 'number line' representation of the numerosities that emerges in the embedding space of the network extrapolates beyond the training range of the network. However, the authors consider only the case of the extrapolating to larger unseen numerosities (e.g., embedding of the samples in range 1-30 by the network trained on numerosities up to 3). One may be curious whether the network could generalize this monotonical line-like representation to smaller unseen numerosities. In Figure 23b one can observe that the model is able to extrapolate to smaller numerosities that the ones have been observed during training. Interestingly, clusters of the numerosities not observed during training (0-10) do not overlap: the representation akin to 'subitizing range' emerges. Moreover, even training with numerosities in range $10-12$ was enough to extrapolate monotonical organization of embedding to unseen smaller and larger numerosities (see Figure 24a). Though, the representation of consequtive numerosities is not disentangled, as corresponding clusters heavily overlap (see Figure 24b).

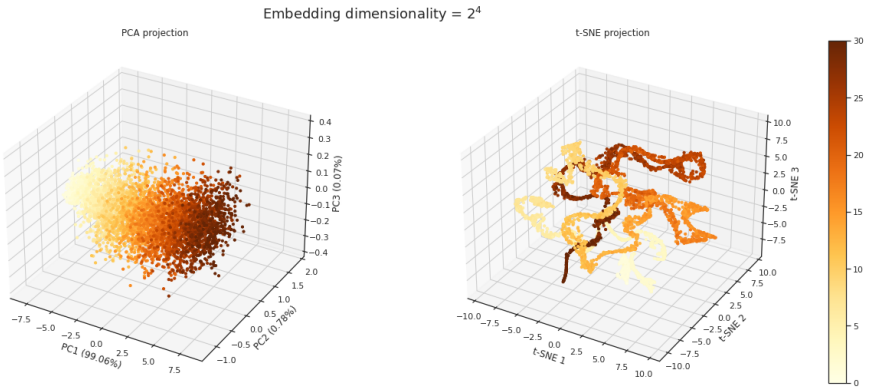
4.1.5 The network architecture is biased to map samples on the line

In [KP20] the authors refer to linear monotonical organization of the embedding space of the network as an emergent representation, that appears as a result of training. We investigated whether the linearity of the embedding is indeed a property that emerges during training. Surprisingly, the encoder is biased to map samples to linear representation (see Figure 25), though samples of different numerosities are initially not organized

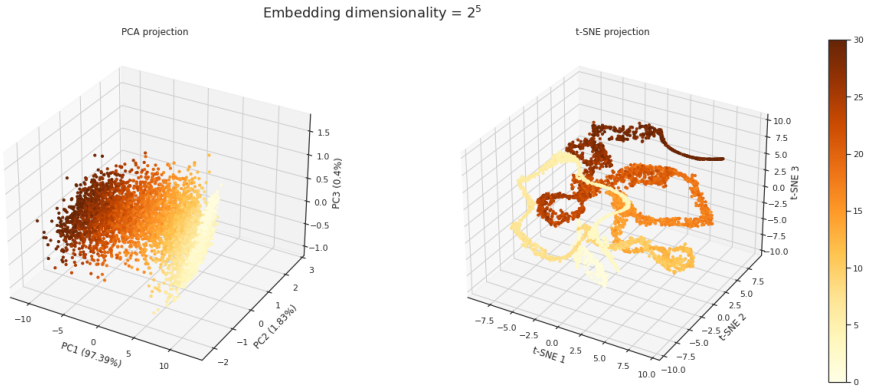
monotonically. This result implies that monotonical organization of embedding space is an emergent property of the training paradigm, while linearity of the learned embedding may be an intrinsic bias of the architecture.



(a)

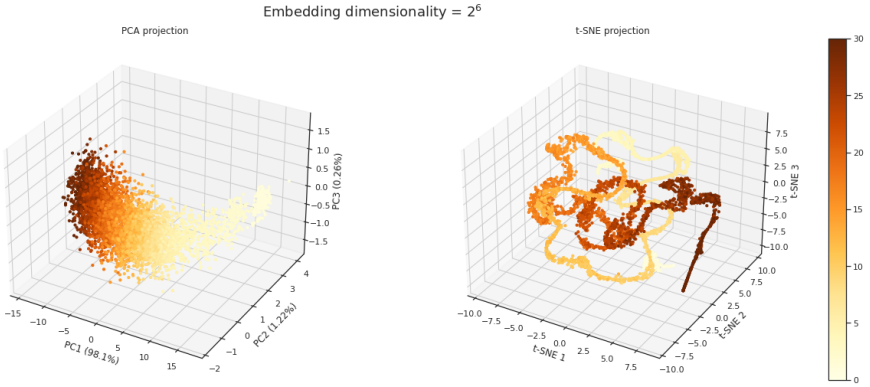


(b)

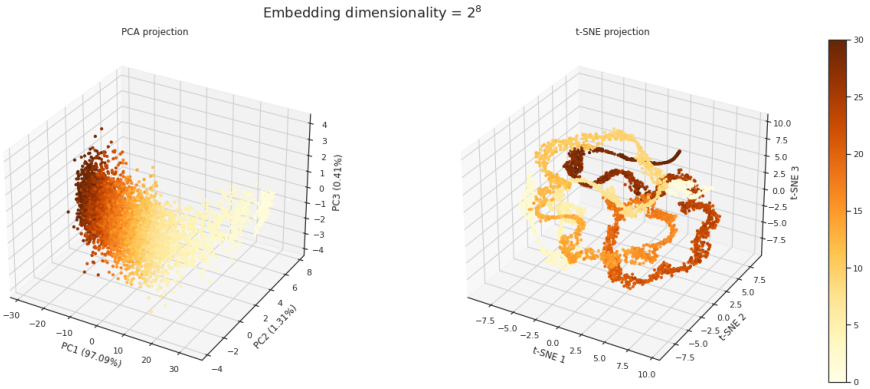


(c)

Figure 21. Visualization of projection of high-dimensional embeddings. The model has been trained on Variable Size, Shape and Contrast dataset with numerosities up to 3 present during training. The embedding visualization is for corresponding *test set* containing numerosities up to 30.



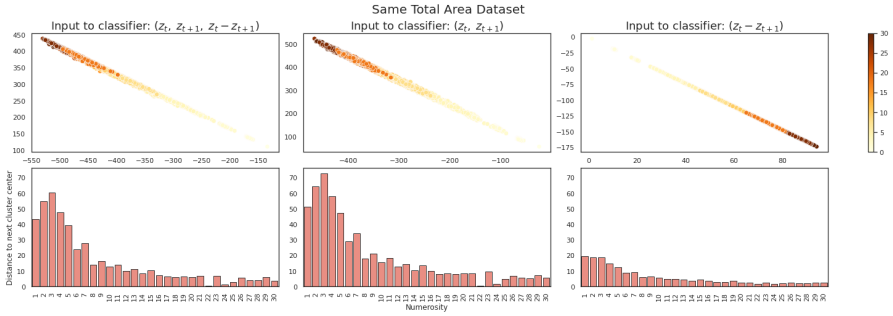
(d)



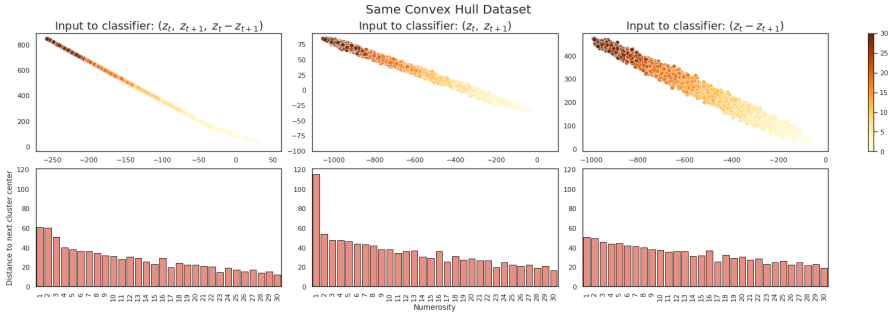
(e)

Figure 21. Visualization of projection of high-dimensional embeddings. The model has been trained on Variable Size, Shape and Contrast dataset with numerosities up to 3 present during training. The embedding visualization is for corresponding *test set* containing numerosities up to 30. (cont.)

(a)

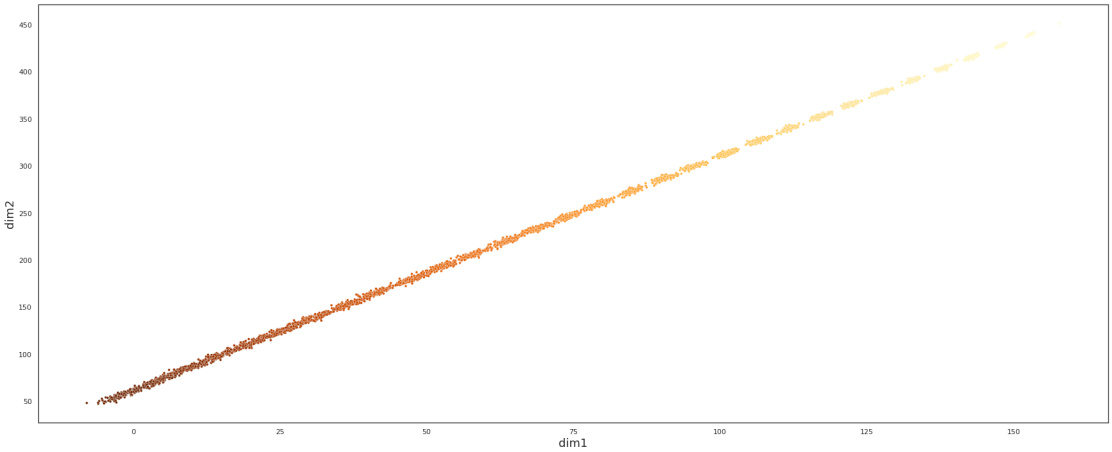


(b)

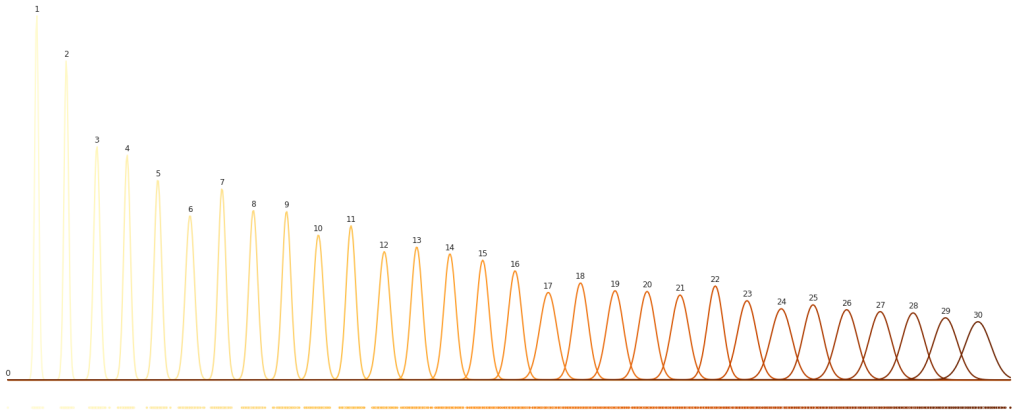


(c)

Figure 22. Effect of the input to the classifier on the embedding structure. For each dataset, the model has been trained with numerosities up to 3 present during training. (a), (b), (c): top row: the embedding visualization is for corresponding *test sets* containing samples of numerosities up to 30; bottom row: distance between centers of consecutive clusters. Cluster center is computed as an average of embedding coordinates of samples of corresponding numerosity.

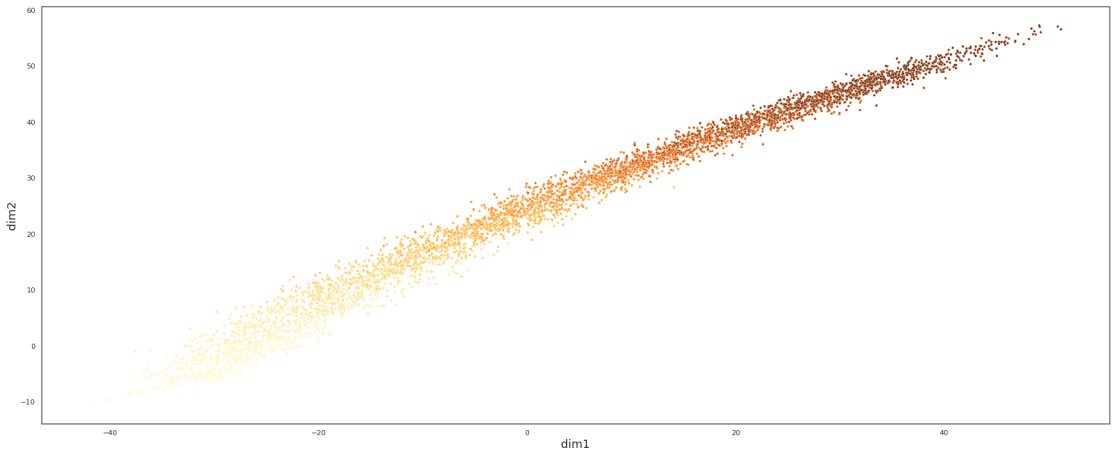


(a) Embedding visualization

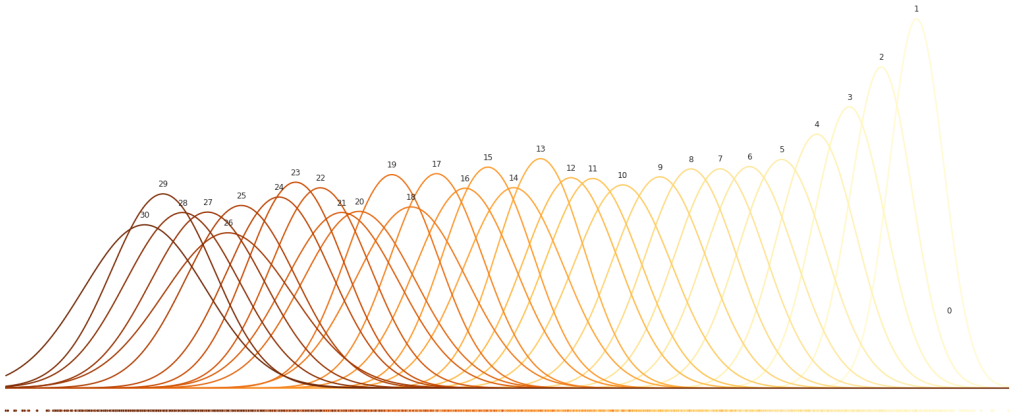


(b) Gaussian fit

Figure 23. Visualization of 'negative' extrapolation: extrapolation to numerosities smaller, than the ones observed during training. The model has been trained on the Standard dataset with numerosities in range 11-30. (a) Embedding visualization of samples from the *test set* with numerosities 0-30. (b) Gaussian fit to the projection on the first principle component (explained variance: 0.99994637) of samples of each numerosity.

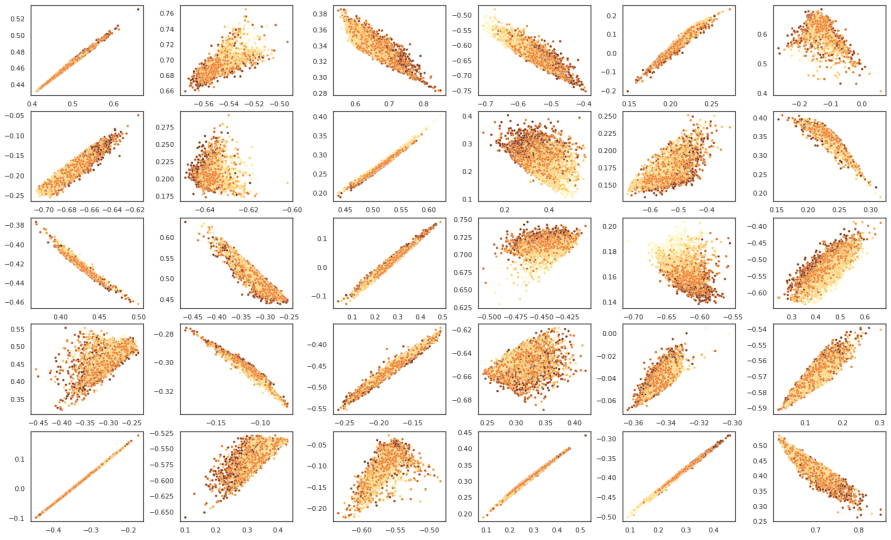


(a) Embedding visualization

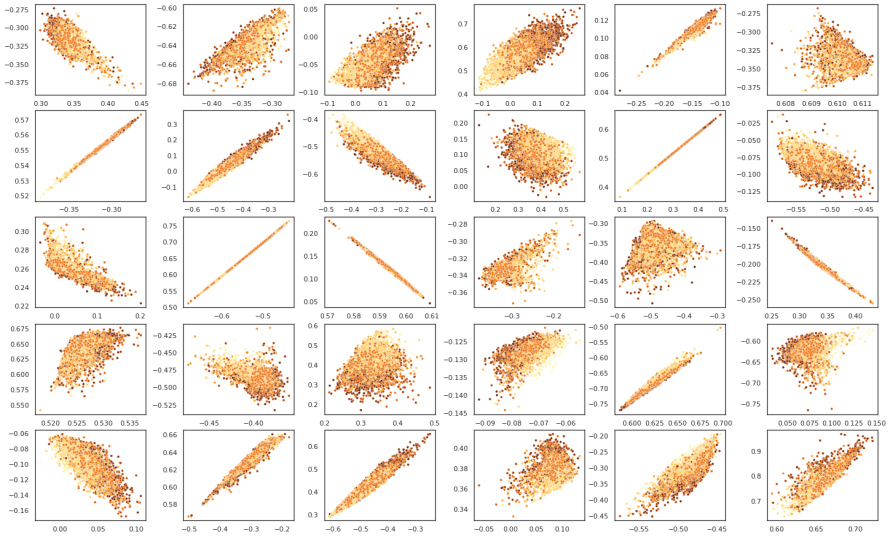


(b) Gaussian fit

Figure 24. Visualization of extrapolation in both directions. The model has been trained on the Standard dataset with numerosities in range 10-12. (a) Embedding visualization of samples from the *test set* with numerosities 0-30. (b) Gaussian fit to the projection on the first principle component (explained variance: 0.99498508) of samples of each numerosity.



(a) first 3 layers are initialized with weights from AlexNet pre-trained on ImageNet, the rest of the weight are randomly initialized



(b) fully randomly initialized extractor

Figure 25. Visualization of embeddings of Variable size and shape dataset by 30 untrained models.

4.2 Results for the detection of numerosity-sensitive neurons in randomly initialized neural network

4.2.1 Numerosity-selective neurons in untrained neural network

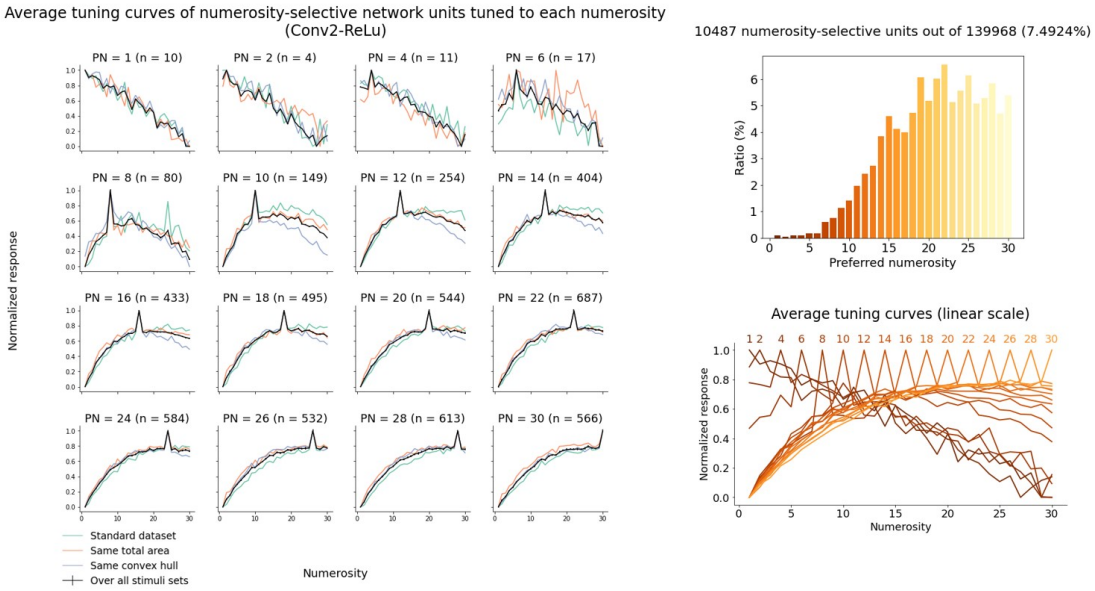
We have replicated the detection of numerosity-selective neurons in randomly initialized AlexNet (see Figure 26). Tuning profiles of units detected in deeper layers of the architecture Conv4-ReLU and Conv5-ReLU resemble the tuning profiles reported in the original work of [KJB$^+$21] (see Figure 3a): average tuning curves of units preferring smaller numerosities are sharper, while as the preferred numerosity of the units increases, the width of the tuning profile increases. The shape of the histogram of distribution of preferred numerosity across number-selective units in layer Conv5-ReLU is similar to the one reported in the original paper (see Figure 3b). In addition to Conv5-ReLU layer analysed in the original work, we have also detected numerosity-selective units in preceding convolutional layers Conv2-ReLU, Conv3-ReLU, and Conv4-ReLU. We observe that the ratio of detected numerosity-selective units to the total number of units in the layer decreases with depth: $\approx 7\%$ in Conv2-ReLU, $\approx 4\%$ in Conv3-ReLU, $\approx 3\%$ in Conv4-ReLU, and $\approx 2\%$ in Conv5-ReLU. Tuning properties of detected numerosity-selective network units change with increasing depth of the layer. Interestingly, the majority of number-selective units in Conv2-ReLU have larger preferred numerosities, while the majority of units detected in the following Conv3-ReLU layer have smaller preferred numerosities.

4.2.2 Additional control for confound variables drastically decreases the number of detected numerosity-selective neurons

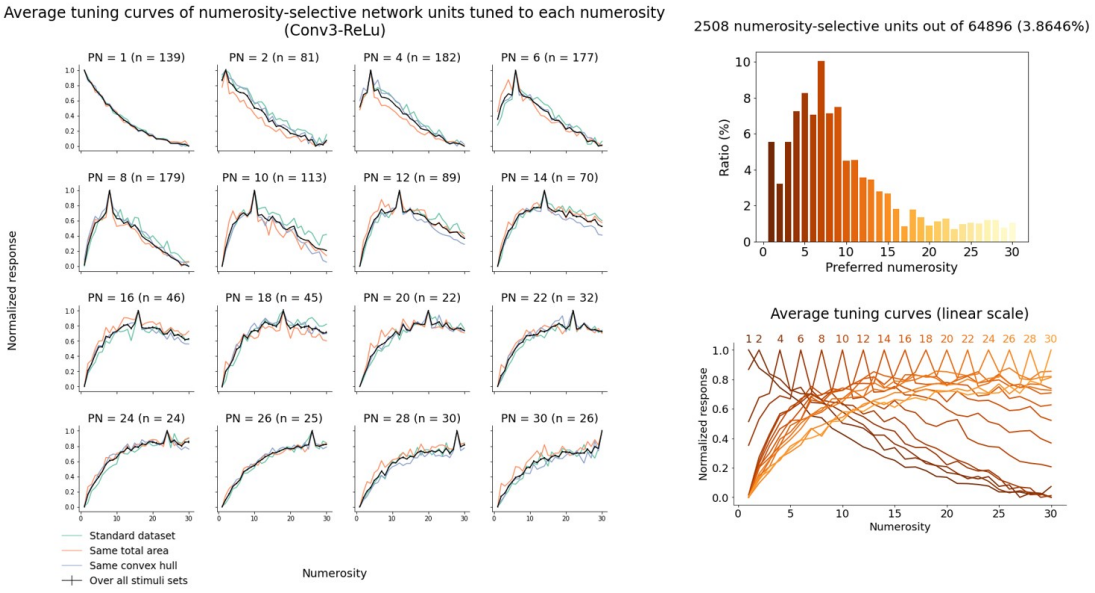
To test how robust are detected neurons with respect to additional control for confound variables, we generated another control stimulus set with control for total perimeter length. Adding this control set as another degree of variability of stimulus set factor in two-way ANOVA analysis **drastically** decreased the number of previously detected numerosity-sensitive neurons in each convolutional layer (see Figure 27): $\approx 63\%$, $\approx 56\%$, $\approx 47\%$, $\approx 38\%$ of number of previously detected numerosity-sensitive units in Conv2-ReLU, Conv3-ReLU, Conv4-ReLU, Conv5-ReLU layers respectively remain numerosity-selective under additional control.

4.2.3 Robustness with respect to morphing deformation

We have investigated how robust are the detected number-sensitive units to morphing deformations. We have generated a Morphing Ellipse dataset in which we gradually morphing a circle into ellipse with increasing eccentricity. If detected units were abstracting numerosity information, we would expect their activity to be invariant with respect to such deformation, since it does not change the numerosity. However, as one can observe



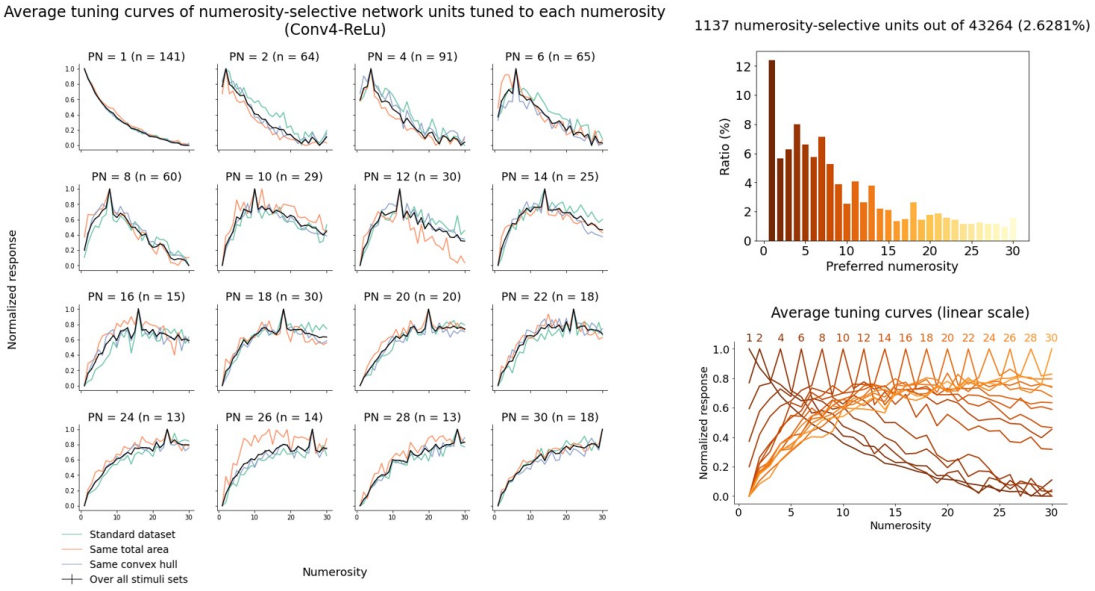
(a) Conv2-ReLU



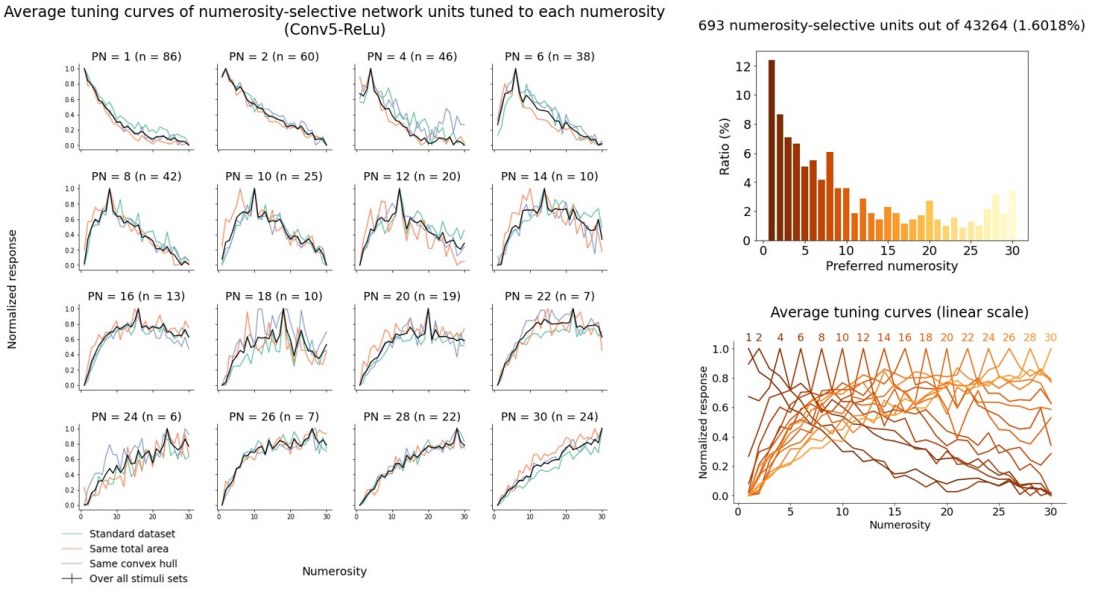
(b) Conv3-ReLU

Figure 26. Visualization of detected numerosity-selective neurons.

in Figure 28, units with $PN = 1$ decrease their activity as the ellipse becomes more elongated, while units with larger preferred numerosities increase their activity.

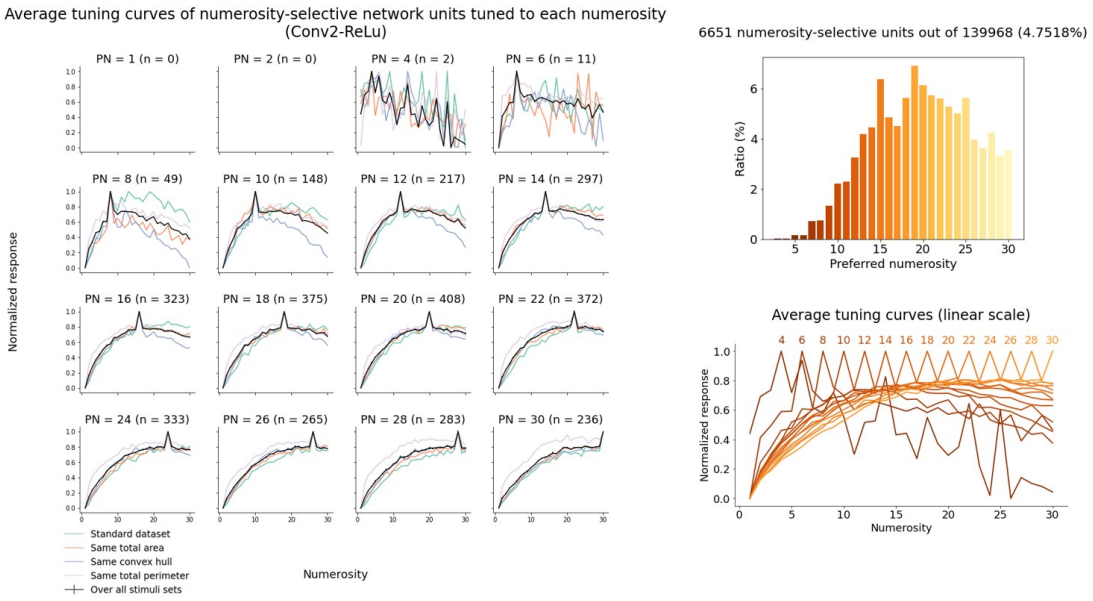


(c) Conv4-ReLU

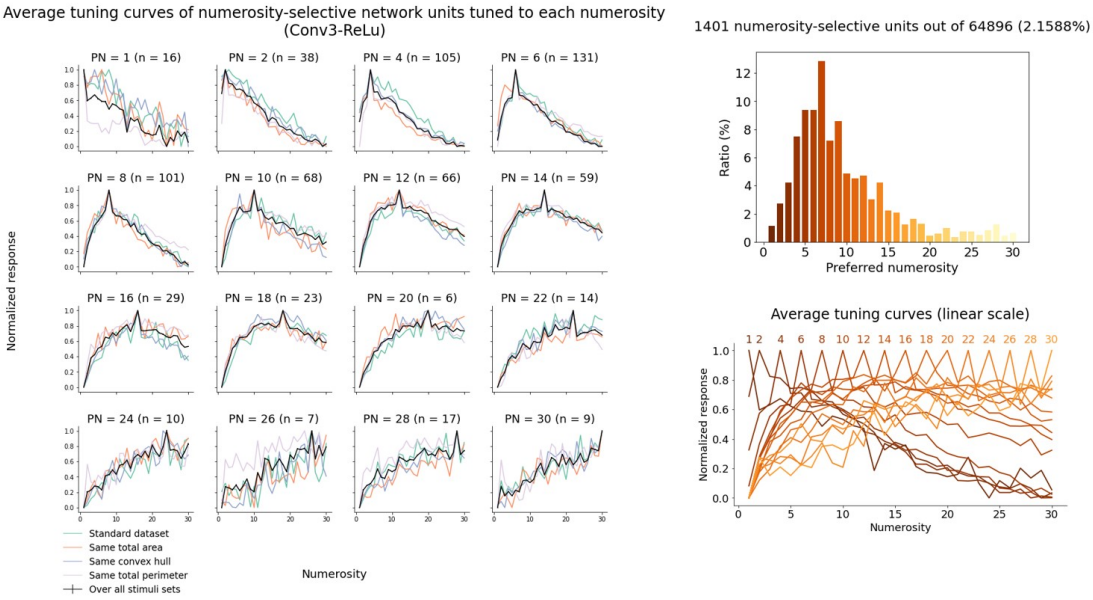


(d) Conv5-ReLU

Figure 26. Visualization of detected numerosity-selective neurons. (cont.)

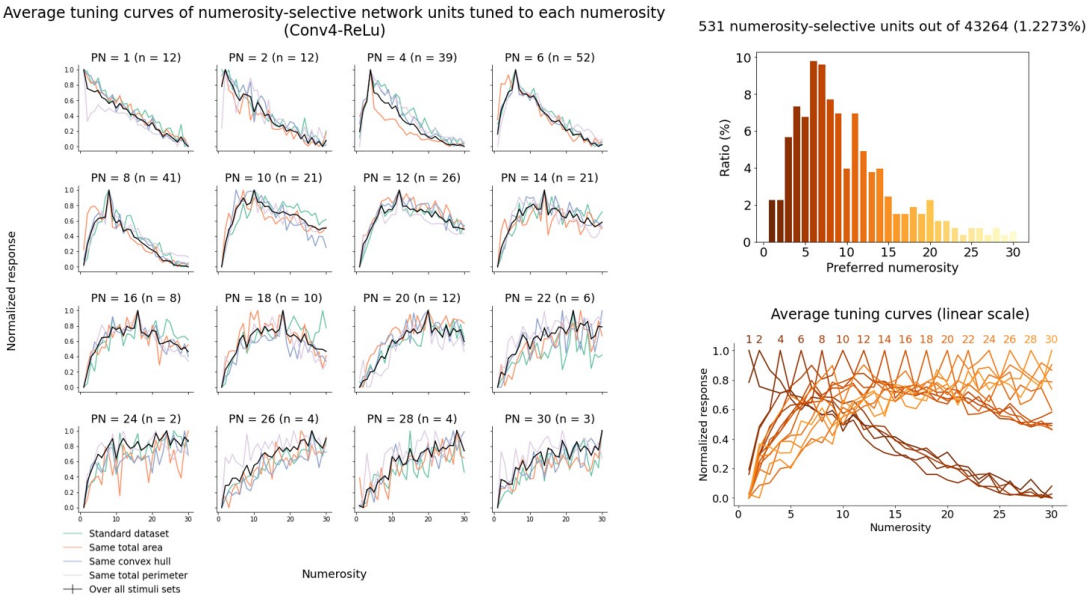


(a) Conv2-ReLU

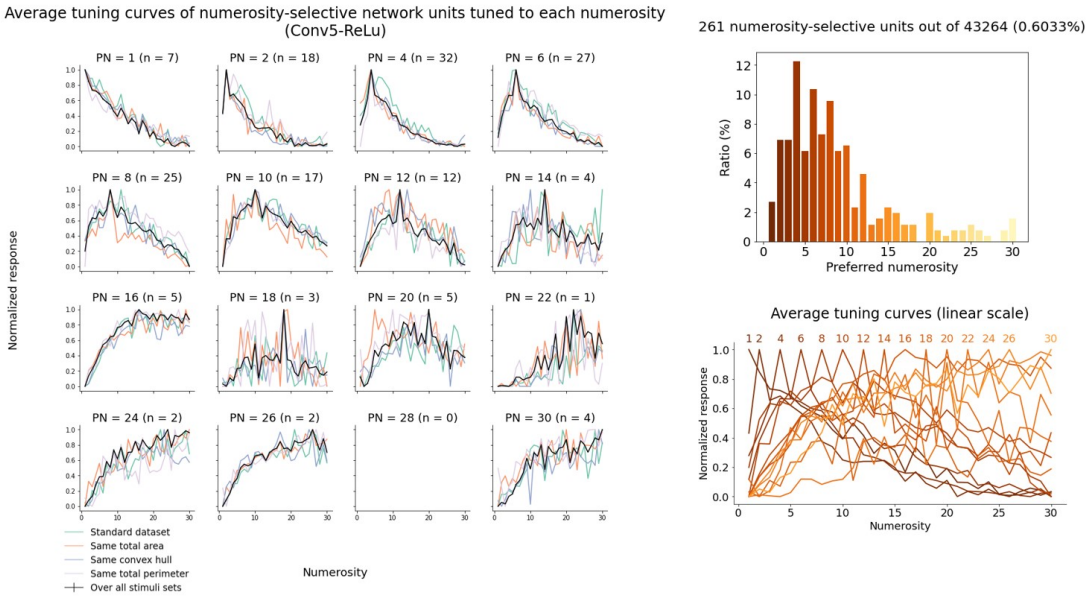


(b) Conv3-ReLU

Figure 27. Visualization of detected numerosity-selective neurons after adding additional control for confound variable (control for total perimeter length).

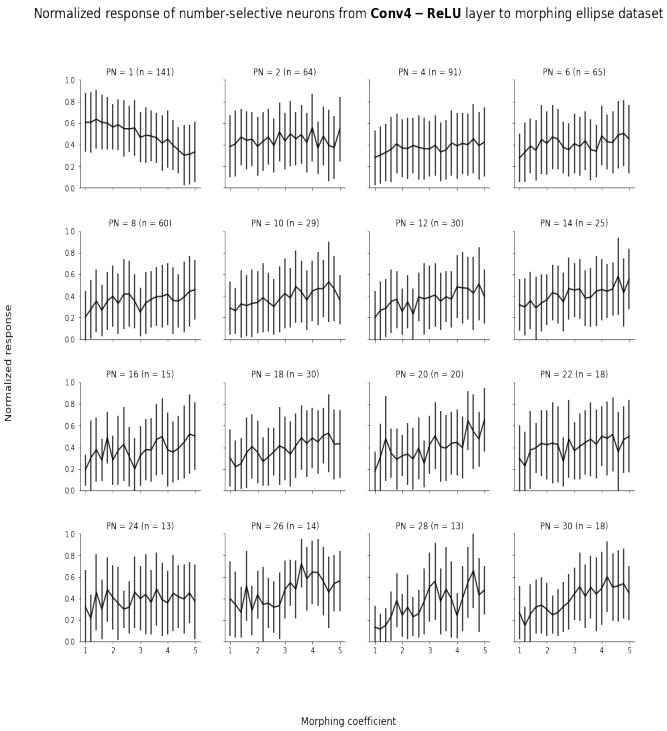


(c) Conv4-ReLU

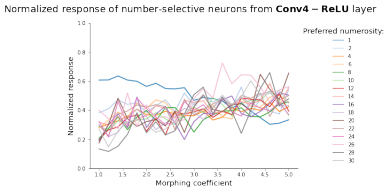


(d) Conv5-ReLU

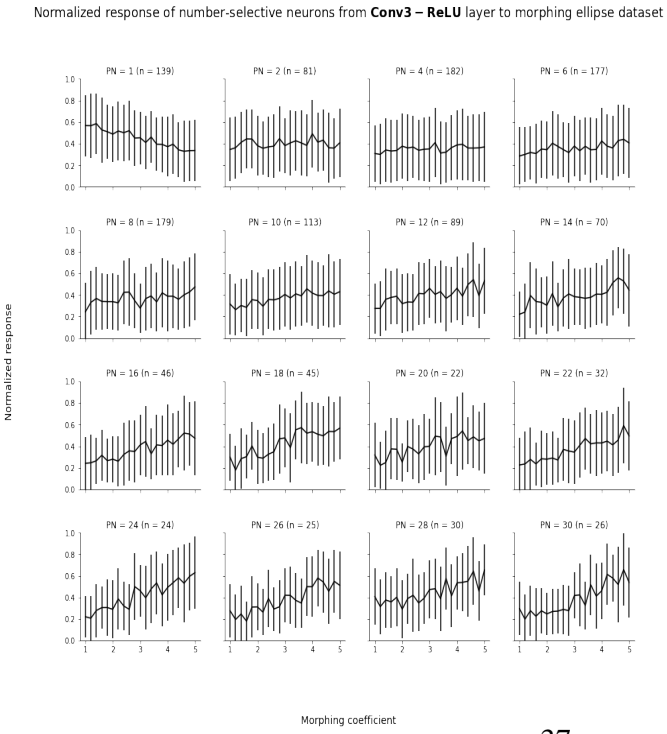
Figure 27. Visualization of detected numerosity-selective neurons after adding additional control for confound variable (control for total perimeter length). (cont.)



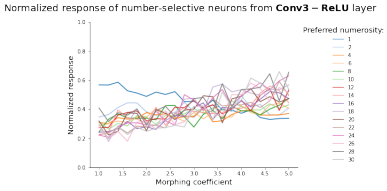
(a)



(b)



(c)



(d)

Figure 28. Response of number-selective neurons to Morphing Ellipse dataset.

5 Discussion

We investigated the robustness of learned representation in the embedding space of manipulative brain model with respect to the visual properties of stimulus object set. We observed that the monotonical organization of the embedding space is extrapolated to unseen numerosities in both directions, and persists in higher dimensions. However, the special topological structure of embedding space is not an abstracted representation of the cardinality of the set, and depends on the variability of the training set. The training paradigm is implicitly akin to contrastive learning paradigm ([CKNH20]): *shake* action provides the model with the information on what degrees of variability of the training samples do not change the cardinality, hence samples of the same numerosity are mapped closer in the embedding space, while *put* and *take* actions mark corresponding samples as "negative examples". However, as we have demonstrated in our experiments, if the degree of variability with respect to which numerosity is invariant (e.g. samples of inverted colors) was not present in the training set, the model fails to embed corresponding "adversarial" sample to the cluster of correct numerosity.
In our results on detection of numerosity-selective neurons in untrained convolutional neural network, we find it surprising that we detected numerosity-selective neurons as early as Conv2-ReLU, since the the receptive field of those units is smaller than the one of units in deeper layers, while the numerosity is a global property of a visual scene. Our further tests of robustness of detected numerosity-selective units with respect to spurious correlations with cardinality demonstrate that their activity is susceptible to changes in total perimeter length. Though, our observation of the connection between total contour length and numerosity estimation goes in line with a recent work of [TDRZ20], report that total contour length significantly impacted the responses of Deep Belief Networks and human participants in a numerosity discrimination task. Our finding also implies that creating a more rigid definition of numerosity-selective neurons would be an important contribution to the field. We also observe that number-selective units were susceptible to topological transformations of the objects that do not change the cardinality, such as stretching. Viewing numerosity perception as a topological task could be one of the steps towards a more rigid formalization of "number neurons", as units which activity is significantly modulated by changes in numerosity, and does not change with respect to topological transformations that leave the cardinality invariant. This is already an ongoing research direction [KZ16].

5.1 Limitations of our work and future work

In our work we have tried to probe the embedding space of the manipulative brain model with "naïve" adversarial examples that demonstrate what impacts the representation learned in the embedding space. However, as the training dataset becomes more and more diverse, it would be harder to probe it by manually coming up with potential

adversarial examples. In our future work, we would like to apply a more rigid analysis to the representations of number neurons in the embedding space of [KP20], as well as numerosity-selective neurons reported in [KJB$^+$21] by generating a stimulus that would maximize the activity of those units via *activation maximization* technique ([EBCV09], [SVZ13], [YCN$^+$15]) that generates the input stimuli that maximizes the response of a given artificial unit by gradient ascent in the input space. We have made initial steps in this direction by making use of implementation available GitHub [NH20] (see Figure 29). However, even though standards regularization techniques like L_2 regularization Gaussian Blur, and others [YCN$^+$15] have been applied during optimization process, stimuli generated this way is quite noisy. In the future, we would like to use method proposed in [NDY$^+$16] which incorporates learned prior on the space on which the gradient ascent is performed, so that the generated stimuli are realistic samples from desired data distribution.

In our replication of detection of numerosity-selective neurons in AlexNet network,

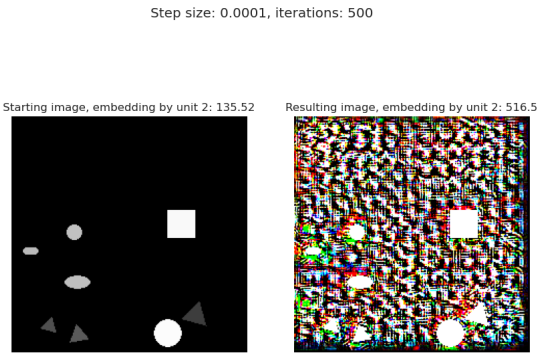


Figure 29. Example of a naive gradient ascent in the input space to maximize an activity of one of units in the final layer of the encoder part of manipulative brain model.

the ratio of detected numerosity-selective units to total number of units in Conv5-ReLU layer of AlexNet was substantially smaller than the one reported in [KJB$^+$21] ($\approx 1.6\%$ vs 8.52% reported in the original paper). We hypothesise that there are two potential causes of this discrepancy. Firstly, the experiments in the original paper were done using MATLAB (MathWorks Inc.) software, while we have used Python libraries and PyTorch framework in our experiments. Secondly, in the original paper the numerosities present in stimulus sets are $1, 2, 4, ..., 30$, hence numerosity factor has 16 degrees of freedom, while we generate samples of all numerosities in range $1, 2, 3, ..., 30$, hence numerosity factor in our analysis has 30 degrees of freedom.

6 Conclusion

In this work we have investigated the ability of artificial neural networks to represent cardinality of the set of objects. Cardinality of a set is an abstract property, independent of properties of the objects that compose this set, hence its representation requires a high level of generalization. Our work is built upon models of numerosity perception in artificial neural networks proposed in [KP20] and [KJB$^+$21]. The first work attributes number sense to the representation learned by the network as a result of implicit contrastive learning. In contrast, the second work reports number sense in untrained neural network, and attributes the emergence of numerosity-selective neurons to the hierarchical structure of the model architecture. In this thesis we have replicated aforementioned models and run additional analysis to test the degree of abstraction of those representations. Our results indicate that in both models the representation of numerosity is not fully disentangled from the representation of continuous visual properties of the objects present in the stimulus.

References

[BAA17] David Burr, Giovanni Anobile, and Roberto Arrighi. Psychophysical evidence for the number sense. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373:20170045, 02 2017.

[BBB+93] Jane Bromley, James Bentz, Leon Bottou, Isabelle Guyon, Yann Lecun, Cliff Moore, Eduard Sackinger, and Rookpak Shah. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7:25, 08 1993.

[BC21a] BoCtrl-C. Non-uniform dots dataset. `https://github.com/BoCtrl-C/lea-nr-ing/blob/main/nud_generation.py`, 2021.

[BC21b] BoCtrl-C. Uniform dots dataset. `https://github.com/BoCtrl-C/lea-nr-ing/blob/main/ud_generation.py`, 2021.

[BR07] David Burr and John Ross. A visual sense of number. *Nature Precedings*, 2, 11 2007.

[BTZ21] Tommaso Boccato, Alberto Testolin, and Marco Zorzi. Learning numerosity representations with transformers: Number generation tasks and out-of-distribution generalization. *Entropy*, 23, 2021.

[But22] B. Butterworth. *Can Fish Count?* Quercus, 2022.

[CKNH20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[DC93] Stanislas Dehaene and Jean-Pierre Changeux. Development of elementary numerical abilities: A neuronal model. *Journal of cognitive neuroscience*, 5:390–407, 10 1993.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[DN15] Helen M. Ditz and Andreas Nieder. Neurons selective to the number of visual items in the corvid songbird endbrain. *Proceedings of the National Academy of Sciences*, 112(25):7827–7832, 2015.

[EBCV09] Dumitru Erhan, Y. Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *Technical Report, Univeristé de Montréal*, 01 2009.

[HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015.

[ID08] Véronique Izard and Stanislas Dehaene. Calibrating the mental number line. *Cognition*, 106(3):1221–1247, 2008.

[KB14] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[KJB+21] Gwangsu Kim, Jaeson Jang, Seungdae Baek, Min Song, and Se-Bum Paik. Visual number sense in untrained deep neural networks. *Science Advances*, 7(1):eabd6127, 2021.

[KP20] Neehar Kondapaneni and Pietro Perona. A number sense as an emergent property of the manipulating brain, 2020.

[KRK14] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLOS Computational Biology*, 10(11):1–29, 11 2014.

[KSH17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2017.

[KZ16] Tobias Kluth and Christoph Zetzsche. Numerosity as a topological invariant. *Journal of vision*, 16(3):30–30, 2016.

[LB95] Yann Lecun and Yoshua Bengio. *Convolutional networks for images, speech, and time-series*. MIT Press, 1995.

[NDY+16] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, 2016.

[NH20] Nguyen-Hoa. Activation maximization. `https://github.com/Nguyen-Hoa/Activation-Maximization`, 2020.

[NM07] Andreas Nieder and Katharina Merten. A labeled-line code for small and large numerosities in the monkey prefrontal cortex. *Journal of Neuroscience*, 27(22):5986–5993, 2007.

[NVN19] Khaled Nasr, Pooja Viswanathan, and Andreas Nieder. Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Science Advances*, 5(5):eaav7903, 2019.

[PGM+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[SP10] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[SVZ13] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[SZ12] Ivilin Stoianov and Marco Zorzi. Emergence of a "visual number sense" in hierarchical generative models. *Nature neuroscience*, 15:194–6, 02 2012.

[TDRZ20] Alberto Testolin, Serena Dolfi, Mathijs Rochus, and Marco Zorzi. Visual sense of number vs. sense of magnitude in humans and machines. *Scientific Reports*, 10, 06 2020.

[TMRP70] Richard F. Thompson, Kathleen S. Mayers, Richard T. Robertson, and Charlotte J. Patterson. Number coding in association cortex of the cat. *Science*, 168(3928):271–273, 1970.

[VF04] Tom Verguts and Wim Fias. Representation of number in animals and humans: A neural model. *Journal of cognitive neuroscience*, 16:1493–504, 12 2004.

[XSG05] Fei Xu, Elizabeth Spelke, and Sydney Goddard. Number sense in human infants. *Developmental science*, 8:88–101, 01 2005.

[YCN+15] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization, 2015.

Appendix

I. Additional visualizations of robustness of number line with respect to dataset variability

Following the analysis done in [KP20], we have studied the behaviour of the network when visual property of the dataset that is confound with numerosity anticorrelates with it. The more variability we encode into the dataset, the more robust it is with respect to the impact of the confound variables. In Figure 31 we observe that the accuracy of the comparative estimation of quantity (see Figure 30) decreases faster for the model trained on Standard Dataset, while it is more stable for the model trained on Variable Size and Contrast Dataset. However, when total area is large enough (small enough) for samples of numerosity smaller (larger) than the reference numerosity, the accuracy of comparative estimation of quantity drops.

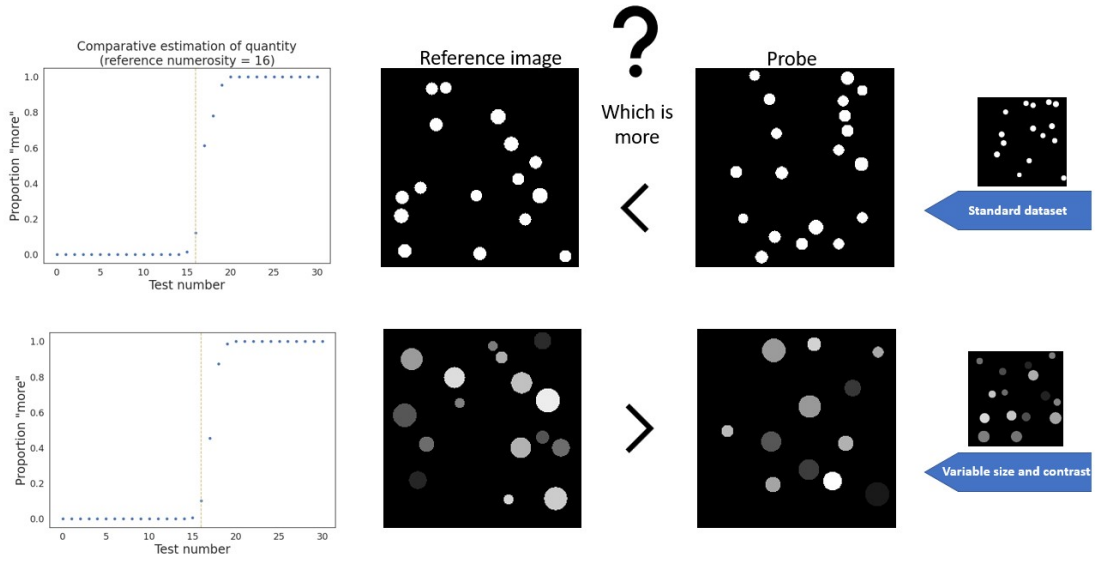


Figure 30. Comparative estimation of quantity. This plot is analog of the psychophysics functions of numerosity comparison in humans [BR07]. The way we construct this plot is we first select a *probe image* with reference numerosity. Next, for each considered numerosity M we have N (here, $N = 150$) test images of this numerosity, and we calculate the proportion of test images of numerosity M that have larger "perceived numerosity" than the probe image. In case of manipulative brain model, the "perceived numerosity" is the distance in embedding space of the embedding of considered sample to embedding of image of zero numerosity. In order to obtain the accuracy of he comparative estimation of quantity, for numerosities larger than the reference numerosity it is "Proportion more" value, while for numerosities smaller than the reference numerosity it is 1 - "Proportion more".

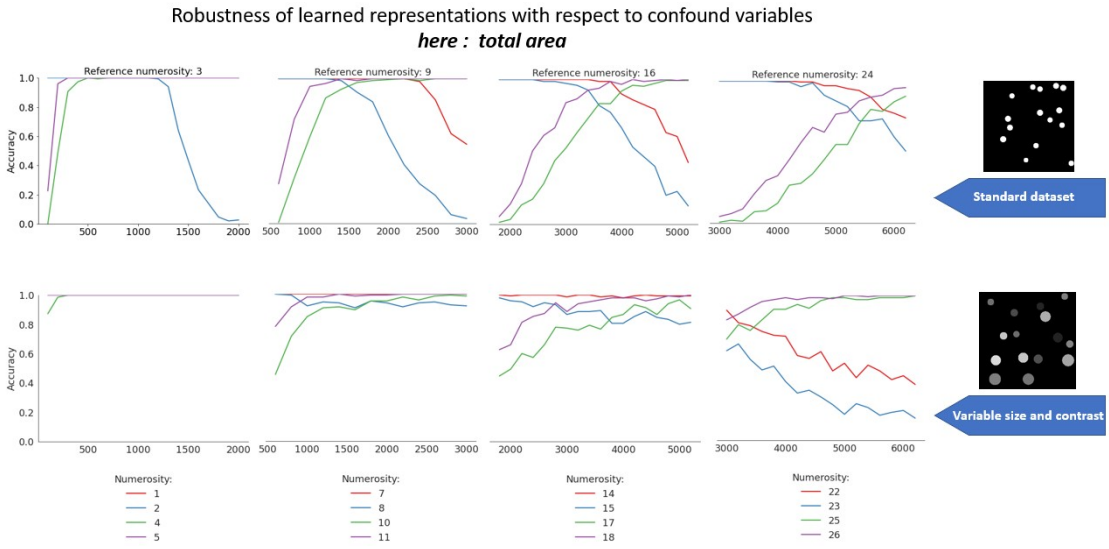


Figure 31. Robustness of the learned representation with respect to change in total area. The accuracy of comparative estimation of quantity for numerosities $i-1,\ i\ , i+1\ , i+2$ vs i for reference numerosities $i \in \{3, 9, 16, 24\}$. Top row: results for model trained on Standard Dataset, bottom row: model trained on Variable size and Contrast dataset. Both models were trained on numerosities up to 3.

II. Access to the Code

The code used to obtain the results can be found in this GitHub repository given below:
`https://github.com/sivomke/numerosity-sense-in-neural-networks`

III. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Tetiana Rabiichuk**,
*(*author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Numerosity Sense in Artificial Neural Networks,
   *(*title of thesis)

   supervised by Raul Vicente Zafra.
   *(*supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Tetiana Rabiichuk
08/08/2022