A Recurrent Model of Approximate Enumeration

Stanford CS238 Final Project

Gabe Barney
Symbolic Systems
Stanford University
barneyga@stanford.edu

Griffin Young
Symbolic Systems
Stanford University
gcyoung@stanford.edu

Abstract

Many animals have the ability to subitize–that is, to rapidly and exactly assess the number of items in a small set (typically less than 5). This has traditionally been considered a separate ability from the Approximate Number System, which allows for approximate estimation of the number of items in a set. Recently, a unified theoretical model was proposed which accounts for the qualitative change in accuracy and speed above the subitizing threshold through a Bayesian model of information gathering under resource constraints [1]. In this paper, we provide a concrete implementation of this theoretical model by using a modified version of the Recurrent Attention Model (RAM), a recurrent neural network which can direct its attention to different points in an image across several steps of processing. We find that, as predicted by the Bayesian model, underestimation occurs in an imbalanced dataset in lower-capacity models and at earlier timesteps.

1 Introduction

Numerical cognition is a field of neuroscience that is being transformed by the use of connectionist models for modeling the brain. One foundational question in this field is the basis of the Approximate Number System. This is the ability of humans and animals to estimate the number of objects in the field of view. In 2012, researchers found neurons in a Deep Neural Network trained only to reconstruct an input image whose responses had a monotonic relationship with numerosity [2], mirroring the response profiles of neurons in the Lateral Intraparietal Area of macaques [3].

More recently, Convolutional Neural Networks pretrained on ImageNet but naive to explicit numerosity training have been shown to have peaked responses to particular numerosities and a graded decrease in response to adjacent numerosities [4][5]. These findings mirrored the tuning properties of neurons in the Ventral Intraparietal Area of nonhuman primates [6] and humans [7].

The temporal characteristics of numerosity processing have not received as much modeling attention. Previous models, with one exception, have assumed a single parallel extraction of numerosity from an image. The one notable exception being [8], though no analysis of the temporal dynamics of the network were presented and the DRAW model allows the model to dynamically alter the shape of the glimpse it takes, unlike biological vision. This is a notable gap in the literature considering the evidence that human numerosity estimates become more accurate over time [9] and with the number of dots foveated [10].

In this work, we model human approximate number perception with a recurrent neural network which chooses, at each time step, a location in the image at which to center its fovea.

1.1 A Unified Account of Numerosity Perception

The theoretical framework we will use to interpret the behavior of the network comes from a paper by Cheyette and Piantadosi (2020) [1]. In it, they accounted for several key psychophysical features of the approximate number system with a Bayesian framework, in which a prior over possible numerosities is updated over time at a certain information gathering rate up to a certain information bound. Specifically, the distribution over the predicted numerosity k given the actual numerosity n, Q(k|n), is chosen to minimize the expected divergence from the veridical numerosity, bounded by divergence from a prior over the probability of encountering a given numerosity P(n). This allows the Approximate Number System and subitizing to be explained by a single mechanism: subitizing just occurs for numbers who fall within the information bound of the prior.

1.2 Goals

We will operationalize the information bound as the size of the hidden layer of the RNN. We will train several models with different hidden layer sizes and test the following hypotheses, pulled from [1]:

1.2.1 Hypothesis 1: Scalar Variability: Larger sets exhibit larger variability in numerosity estimates.

This is a standard psychophysical finding in humans and animals. Larger set sizes will have more variable numerosity estimates.

1.2.2 Hypothesis 2: Subitizing: Rapid and exact enumeration of small sets.

Small numerosities will be enumerated with very low error and relatively quickly due to the low KL divergence between the power law distribution of numerical frequency (that is, lower numbers are seen more frequently than higher numbers) and the posterior distribution for images with low set sizes.

1.2.3 Hypothesis 3: Subitizing range varies with information bound.

The size of the information bound, operationalized as the size of the hidden layer in the RNN, will determine the cutoff point for the subitizing range. That is, models with larger hidden layers will have a larger range with extremely low error, followed by a transition point to a range of numbers exhibiting scalar variability.

1.2.4 Hypothesis 4: Underestimation of numerosity at lower information bounds.

Again due to the power law distribution, we expect underestimation at lower information bounds.

1.2.5 Hypothesis 5: Worse, and underestimated, predictions at earlier time steps.

This is for the same reason as for the lower information bounds, since the constraint is the minimum of the information bound and the value Rt, where R is a linear information extraction rate and t is the number of time steps.

2 Related Work

2.1 DRAW: A Recurrent Neural Network For Image Generation

The model that has inspired the use of foveated input is the *Deep Recurrent Attentive Writer* (DRAW)[11] architecture. The DRAW model was used by DeepMind to incorporate a mechanism that emulates the foveation of the human eye with a sequential variational autoencoder framework to generate complex images iteratively. The temporal nature of RNNs causes this architecture to provide it with the iterative image construction being more life-like than other approaches to image generation.

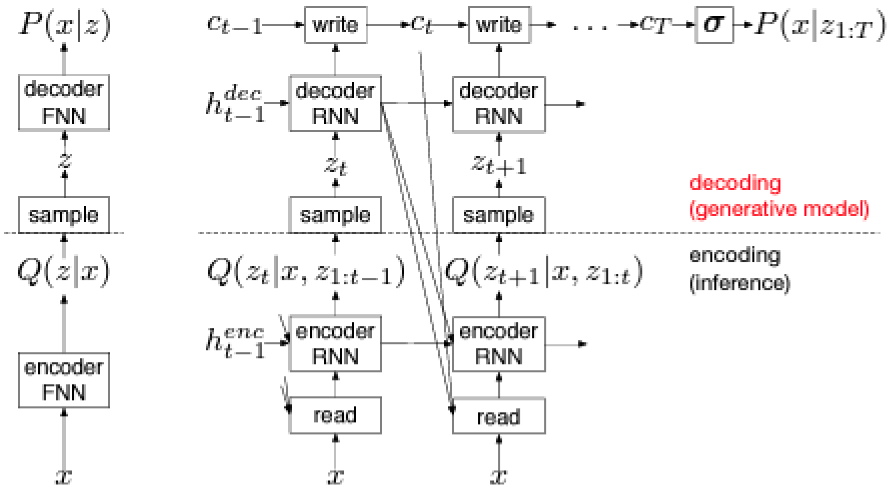


Figure 1: **Left:** Conventional VAE **Right:** DRAW

The architecture of DRAW is a slight alteration to the conventional variational autoencoder. Instead of the encoder and decoder being feedforword neural networks, they are LSTM RNNs. These RNNs take input from the previous encoders and decoders by means of a canvas matrix, which can be partially updated rather than completely generated due to selective attention.

Generating images iteratively enables selective attention to parts of the image. The main challenge is directing that selective attention. DRAW applies an array of 2D Gaussian filters to the image, which varies the location, zoom, and resolution of the attention based on the output of the decoder LSTM.

2.2 Can Generic Neural Networks Estimate Numerosity Like Humans?

Chen, et al. (2018) [8] investigated how well the characteristics of experimental data about the approximate number system would emerge in a standard feedforward network and in a modified version of the DRAW model. The experimental data had shown that there is a constant coefficient of variation for numerosities larger than four, even in datasets without imbalanced numerosities. They found that the sequential nature of the model was not used, as "the focus of attention remained constant over glimpses, and the accuracy of estimates was almost as good on the first as on the last glimpse."
In this paper, we wanted to focus moreso on the temporal aspect of glimpses, and determining how time affects the models understanding. , so we chose the RAM model instead of the DRAM model. We discuss DRAM in section 3.5.

2.3 Numerosity discrimination in deep neural networks: Initial competence, developmental refinement and experience statistics

Humans, non-humans, and their babies can all detect changes in numerosity. This is evidence for the innate number system. The possibility for neural network modeling sheds light on potential explanations for why numerical acuity improves with cognitive development. In this paper, Testolin, et al. (2019) [12] substantiates the basis of the innate number system by showing that neural networks exhibit numerosity sensitivity without training and that numerical acuity is improved our time through training. Their main finding was that animals may not necessarily have dedicated systems for numerosity, because their use of a general neural network demonstrates that a general system can provide this processing, rather than an evolutionarily-based approximate number system.

2.4 Recurrent Models of Visual Attention (RAM):

Mnih, et al. (2014) [13] presents the Recurrent Attention Model. Its objective is to recognize objects in images with a deep recurrent neural network via reinforcement learning. At each time step, it received several patches of an input image at different resolutions, centered around some point in the image. It updates its hidden layer and then outputs an output distribution over the target classes as well as a distribution over the next center from which the next patches are sampled.

2.5 Multiple Object Recognition with Visual Attention (DRAM):

Ba, et al. (2015) [14] presents the Deep Recurrent Attention Model. Clearly an alteration of the RAM model by merely inspecting the name. Its objective is to recognize multiple objects in images with a deep recurrent neural network via reinforcement learning, similarly to the RAM model. Its results showed that it was better than the state-of-the-art convolutional neural networks at the time, and used less resources and computation.

RAM is more in line with the domain we chose to look into, as it determines gazing strategies of the images, but finds particular challenges with real world challenges. These numerosity estimation tasks do fall into tasks that RAM is able to model, and so we figured that focusing on gazing and temporal aspects was more important, as well as it is more simplisitic and similar to human behavior.

2.6 Enriched Deep Recurrent Visual Attention Model for Multiple Object Recognition (EDRAM):

A recent paper that we have looked into is the paper that introduces the Enriched Deep Recurrent Visual Attention Model. Ablavatski, et al. (2017) [15] developed an improvement to DRAM and is fully differentiable. The important aspect about this, is the use of the the Spatial Transformer and improved upon the state-of-the-art. Transformers have made a big impact on Natural Language Processing and Computer Vision, so this paper introduces the possibility of applying transformers to this domain.

3 Methods

3.1 Datasets

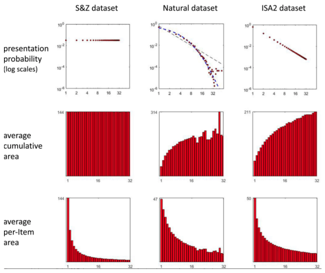


Figure 2

Jay McClelland generously provided the three datasets in figure 2, though we only trained on the Natural and S+Z datasets. The figure summarizes the differences, and full details are available in [12]. Essentially, each contains binary images of varying numbers of white rectangles on a black background. The Natural dataset has a power law distribution, where the frequency of a number n is about $\frac{1}{n^2}$ the frequency of the number 1. The dataset was split into 771840 training samples, 85760 validation samples, and 214400 test samples. The S+Z dataset has a uniform distribution across numbers. We split it into 36864 training samples, 4096 validation samples, and 10240 test samples.

3.2 Model

The model we chose is the Recurrent Attention Model (RAM) from [13]. The core is a recurrent neural network which, at each time step, takes in a foveated 'glimpse' of an input image and outputs: 1) a Gaussian distribution over fovea centers from which the fovea center at the next time step is sampled, 2) a softmax distribution over some classes, and (our novel contribution), 3) a Bernoulli distribution from which a STOP action is sampled, either 1 to continue computation or 0 to halt. The model is trained with the REINFORCE update rule, which adjusts the parameters of the network so that actions which have led to high reward become more probable.

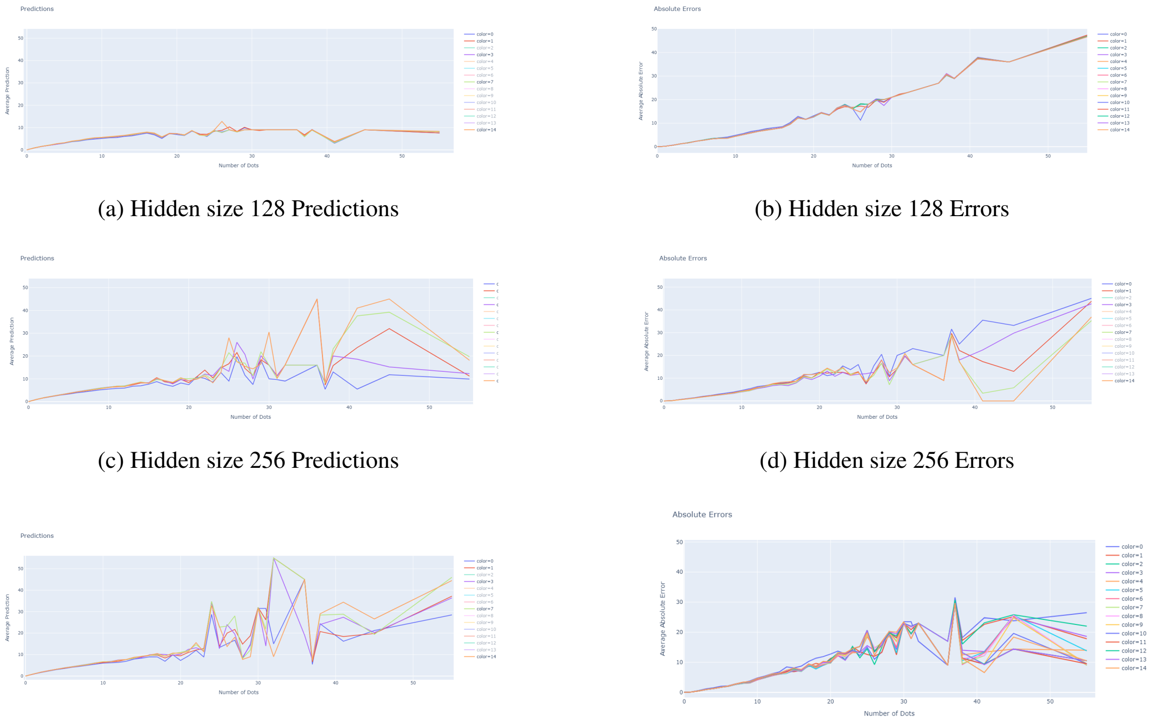
We used an open source RAM implementation [16] as our jumping off point. Instead of giving a reward of 1 to every glimpse step if the final glimpse classified the image correctly, we give a reward of 1 only to the first time step whose STOP action was a 0, and we give a penalty to each time step whose STOP action was a 1.

Our final analysis uses only the original RAM implementation without the STOP action. We found that including any penalty per timestep caused the network to immediately stop computation, not allowing us to examine the evolution of the prediction across time steps. We did include one alteration to the RAM model: instead of only calculating the reward based on the prediction at the final timestep, we calculated the reward based on the predictions at every time step. This incentivized the model to output as accurate an answer as it could at every time step.

3.3 Training

We trained three models, with hidden layers of 64, 256, and 2560, in Google Colaboratory. Patches of the image, centered around the model-chosen fovea center at each time step, were of size 3x3, 12x12, and 48x48, the latter two average-pooled to be 3x3 as well. Each image was viewed for 15 timesteps. We trained the models for 10, 20, and 10 epochs respectively with an ADAM optimizer with initial learning rate 0.0003. Testing was done on the versions of each model with the highest validation accuracies.

4 Results



(a) Hidden size 128 Predictions

(b) Hidden size 128 Errors

(c) Hidden size 256 Predictions

(d) Hidden size 256 Errors

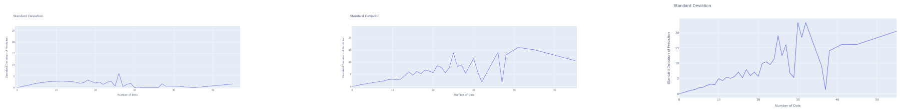
(e) Hidden size 2560 Predictions

(f) Hidden size 2560 Errors

Figure 3: Natural dataset, with power law frequencies across numbers. Predictions and errors averaged across images with the same number of dots at each time step. Each different colored line is a different timestep.

4.1 Hypotheses

- **Hypothesis 1: Scalar variability**
  We found evidence for this hypothesis. For all models but the first, the standard deviation of the predictions increases with the numerosity.



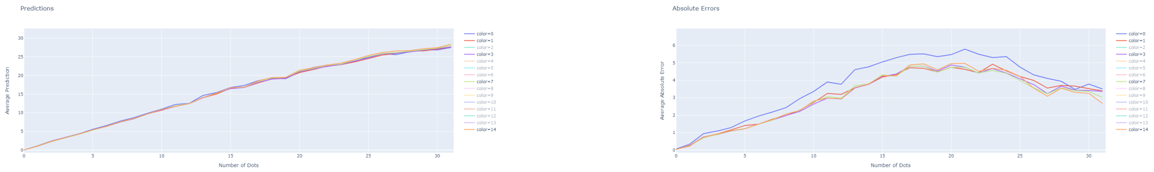
(a) Hidden size 128

(b) Hidden size 256

(c) Hidden size 2560

Figure 4: S+Z dataset, with equal freqeuncies across numbers. Standard deviations of predictions averaged across timesteps for each numerosity.

- **Hypothesis 2: Subitizing**
  We did not find evidence for this phenomenon in our models.
  Without the stop action, we cannot make claims as to the rapidness while within the subitizing range, but we can say that there is stronger confidence of the model in its predictions at lower numerosities. That is, there is less divergence between the predictions between timesteps at lower numbers of dots presented in the images.
  In terms of accuracy, we reproduced previous findings failing to replicate the human-like near-perfect classification accuracy for numbers 2-4 in neural models [8].
- **Hypothesis 3: Subitizing range increases with information bound**
  We did not find support for this hypothesis, since the subitzing range for all the models was the same: only the number 1.
  With a more generous classification of subitizing as anything with less than 1 absolute error, there is a small, yet present, difference in the subitizing ranges between information bounds. Our largest hidden layer sizing has an absolute error of less than 1 at its best timestep until 4 dots are presented, while our smallest size has an absolute error of less than 1 at its best timestep until 3 dots. This is just reflective of a larger trend of lower error in the models with larger hidden layers.



(a) Hidden size 2560 Predictions

(b) Hidden size 2560 Errors

Figure 5: S+Z dataset, with equal freqeuncies across numbers. Predictions and errors averaged across images with the same number of dots at each time step. Each different colored line is a different timestep.

- **Hypothesis 4: Underestimation at lower information bounds**
  We did find evidence to support this hypothesis. There is a progressive decrease in underestimation as the hidden layers size, or information bound, is increased. This is especially notable in contrast to the S+Z dataset in Figure 5, in which the average estimate is actually an overestimation.
- **Hypothesis 5: Underestimation at earlier timesteps**
  We did find evidence to support this hypothesis. We find in the majority of cases on our largest hidden layer model and in every case on our average model that timestep 0 produces the worst performing predictions. Especially in the larger numerosities, we also find a graded increase in predictions as the timesteps progress, with the final timestep having the highest prediction.
  This is especially interesting in contrast to the balanced S+Z dataset in Figure 5, in which,

at the smaller numerosities, timestep 0 is an overestimate compared to the final timestep, but at larger numerosities, timestep 0 is an underestimate. This supports the hypothesis that the model maintains a prior over the probability of each numerosity appearing which it updates over time to be more consistent with the evidence it accumulates. This prior has more mass in the lower numerosities for the Natural dataset, causing underestimation for higher numbers. In the balanced S+Z dataset, however, the prior has equal mass over all numerosities, causing lower numerosities to be overestimated and higher ones to be underestimated.

5 Discussion

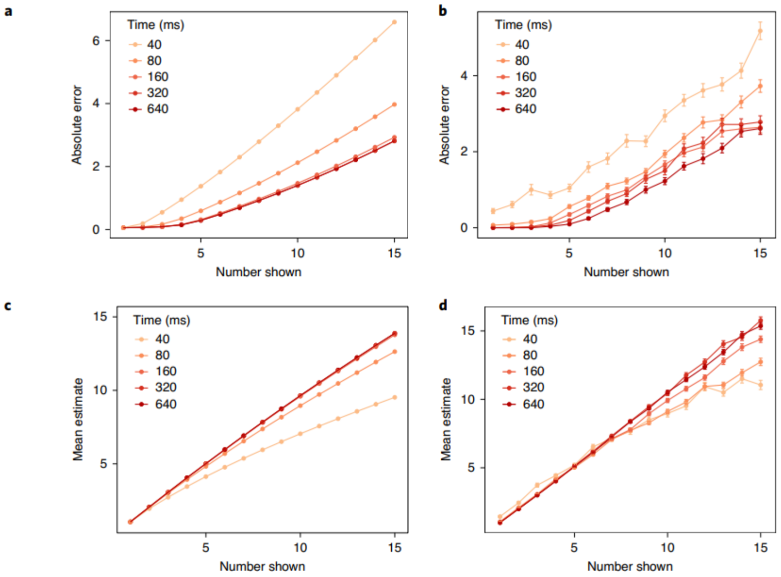


Figure 6: Graph from [1]. The left shows the theoretical model's predictions (top) and error (bottom) for different exposure times. The right shows the same but for human data.

Aside from the subitizing range, our results look very similar to those in Figure 6 for both the theoretical model and the human results. When trained on the imbalanced Natural dataset, our models consistently underestimated the numerosity of the images, especially at earlier timesteps. When trained on a balanced dataset, this pattern shifted to be one of overestimation at lower numbers and underestimation at higher numbers. This is consistent with the unified model presented in [1], in which a prior over numerosities reflecting the frequency of exposure in the dataset is updated over time.

Additionally, by training three models with different hidden layer sizes, we found limited support for the hypothesis that the different subitizing ranges between infants, adults, and chimpanzees are the result of different visual memory capacities. Though none of our models had a subitizing range larger than 1, we did find that the error for both of the larger models was smaller across numerosities. It's possible that an even higher capacity model would push this error down even farther, causing more than just the first number to have near-zero estimation error and thus extending the subitizing range.

Curiously, despite the increase in accuracy across time steps (especially between the first and the second), we found that, as in previous work, the model did not effectively manipulate the location of its retina [8], preferring instead to leave it in mostly the same place in the upper left hand corner of the image. This makes the behavior of the model difficult to compare to humans, who do strategically move their gaze around an image. We were unable to determine why this capability was not used.

Finally, a word about the project of modeling human behavior with deep neural networks:

An issue that we find with our lower hidden layer sizing models is the lack of a notable divergence between timesteps for their predictions. We believe this is because different information bounds reach their capacity for learning this problem at different rates, and the number of epochs they were ran for were either too few or too many. If it were too few, the model may have been able to sufficiently learn from additional timesteps, producing nearly the same results. Or, if it were overtrained, then it may have been able to learn its best numerosity predictions at each timestep, so new timestops do not yield any improvement in performance. This yields some additional discussion.

We question the value of if we should train the RAM model, and if so, how many epochs? These are issues of what is the most effective way to think about RAM's simplistic emulation of human glimpses for this task? It makes sense that numerosity estimation would improve with age [8], and we could think of a fresh RAM model as a newborn child. We do witness some of the hypotheses in models that were trained for a single epoch, but with more epochs the model becomes far more effective in making guesses outside of its subitizing range. But if the models are trained too much, the absolute error throughout this task stays the same no matter the number of dots presented, which does not support the two-system model. It's a very intricate balance, as there is no right or wrong answer for what is the right choice, although it does seem like if we want a reasonable conjugate to human abilities we would have to train the model, but we cannot allow it to become too good at its task, as so it outdoes humans.

6 Conclusions And Future Work

In our concrete model of sequential number estimation, we found evidence to support two key claims made in [1]: that underestimation occurs in systems with lower information bounds and at shorter exposure times due to the imbalance in exposure to different numbers in natural images. Since the subitizing range was the same for all of our models (including an unpublished run with 10,240 neurons in the hidden layer), we were unable to make any conclusions about the effect of visual memory capacity on subitizing range except to say that our higher capacity models did have lower errors across all numerosities, including 2-4, and it's possible that an even higher capacity model would be able to improve performance to the point of extending the subitizing range. This is a promising avenue of further research and would provide compelling evidence for a one-system account of subitizing and approximation.

This Recurrent Attention Model is notably simplistic, and so there are questions as to what changes are necessary to make it a better model of human behavior. There has been many large leaps in modeling since RAM was introduced in 2014, so we may have missed a more effective method, or there is new modeling techniques that could be leveraged into being applied in this domain. Exploring other modeling options could be fruitful, yet we want to maintain the importance of gazing strategies and glimpse duration. Central to this goal is finding a way to train the model to move its retina strategically over time. We are not quite sure if DRAM, EDRAM, or other transformer-based methods can provide us much for this problem. There is certainly much room to explore in the new methods that have even occurred since EDRAM was introduced.

7 Member Contributions

Griffin Young: Finding and adapting the datasets and models. Literature review of psychophysical data and previous numerical cognition studies.

Gabe Barney: Analyzing data. Literature review of similar models and previous numerical cognition studies.

Our source code and high-resolution plots can be found in this public GitHub repo: https://github.com/barneyga/A-Recurrent-Model-of-Approximate-Enumeration

References

[1] Samuel J. Cheyette and Steven T. Piantadosi. A unified account of numerosity perception. *Nature Human Behaviour*, 4(12):1265–1272, December 2020.

[2] Ivilin Stoianov and Marco Zorzi. Emergence of a 'visual number sense' in hierarchical generative models. *Nature Neuroscience*, 15(2):194–196, February 2012.

[3] Jamie D Roitman, Elizabeth M Brannon, and Michael L Platt. Monotonic Coding of Numerosity in Macaque Lateral Intraparietal Area. *PLoS Biology*, 5(8):e208, July 2007.

[4] Nicholas K. DeWind. The number sense is an emergent property of a deep convolutional neural network trained for object recognition. preprint, Animal Behavior and Cognition, April 2019.

[5] 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, May 2019.

[6] A. Nieder and K. Merten. A Labeled-Line Code for Small and Large Numerosities in the Monkey Prefrontal Cortex. *Journal of Neuroscience*, 27(22):5986–5993, May 2007.

[7] Esther F. Kutter, Jan Bostroem, Christian E. Elger, Florian Mormann, and Andreas Nieder. Single Neurons in the Human Brain Encode Numbers. *Neuron*, 100(3):753–761.e4, November 2018.

[8] Sharon Y Chen and Mengting Fang. Can Generic Neural Networks Estimate Numerosity Like Humans? page 6.

[9] Sampling from the mental number line: How are approximate number system representations formed? | Elsevier Enhanced Reader.

[10] Samuel J. Cheyette and Steven T. Piantadosi. A primarily serial, foveal accumulator underlies approximate numerical estimation. *Proceedings of the National Academy of Sciences*, 116(36):17729–17734, September 2019.

[11] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. DRAW: A Recurrent Neural Network For Image Generation. *arXiv:1502.04623 [cs]*, May 2015. arXiv: 1502.04623.

[12] Alberto Testolin, Will Y. Zou, and James L. McClelland. Numerosity discrimination in deep neural networks: Initial competence, developmental refinement and experience statistics. *Developmental Science*, 23(5), 2020.

[13] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention, 2014.

[14] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention, 2015.

[15] Artsiom Ablavatski, Shijian Lu, and Jianfei Cai. Enriched deep recurrent visual attention model for multiple object recognition. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar 2017.

[16] Kevin Zakka. Recurrent visual attention. `https://github.com/kevinzakka/recurrent-visual-attention`, 2020.