Numerosity Understanding of Visual Transformers

Bharat Goyal (bgoyal7@gatech.edu)
College of Computing
Georgia Institute of Technology

Hiren Kumawat (hkumawat3@gatech.edu)
College of Computing
Georgia Institute of Technology

Abstract

Humans are able to visually discriminate numerosities and there has been debate regarding whether such knowledge is learnt or innate. Prior works using CNNs seem to pick up on numerosity without an explicit training objective geared towards the concept itself. Newer models like vision transformers (ViTs) have shown to achieve state-of-the-art performance and stronger understandings of long-range dependencies; experimenting with ViTs' ability to comprehend numerosity is a promising area of research and can provide insight as to whether the human-alignment observed with CNNs is architecture specific. In this work, we use a pre-trained ViT and experiment with its ability to understand various types of numerosity stimuli. We show competitive results against VGG-19 models and even outperform in certain tests. We further analyze our results using attention maps and activation maps.

Keywords: Numerosity; Visual Transformer (ViT); Convolutional Neural Network; Mental Number Line (MNL); Ratio Effect; Activation Map; Attention Map

Introduction

Numerosity, the ability to visually discriminate quantities of items, has long been a subject of inquiry in cognitive science. Humans – both adults and infants – as well as even animals are able to understand the relative magnitude between groupings of items. There has been much debate in the research community as to whether this ability is learned from experience or an innate skill (Spelke & Kinzler, 2006). Notably, recent investigations using Convolutional Neural Networks (CNNs) have demonstrated an ability to discern numerosity, even in the absence of explicit training objectives dedicated to this numerosity concept (Upadhyay, 2023). However, it is also well known that CNNs have an inherent inductive bias in identifying spatial correlation and local structures, which could be a contributor to their success in this task (Wang & Wu, 2023).

Newer models such as Vision Transformers (ViTs) have showcased state-of-the-art performance and a superior understanding of long-range dependencies in visual information processing (Dosovitskiy et al., 2021). The question arises: Can ViTs replicate the human-like numerosity discrimination observed in CNNs, or is this alignment architecture-specific? This research aims to explore the potential of ViTs in comprehending numerosity, shedding light on the intricacies of numerical cognition in artificial intelligence. Additionally, prior research has considered Transformer architectures to be quite close to the human brain functionality. Kozachkov et al. (2023) believe that core Transformer operations can be "naturally implemented" with biological computational units.

In this study, we plan to leverage a pre-trained ViT model due to the model's demonstrated success in capturing complex visual patterns and relationships. Our experimental design will involve exposing the ViT to various numerosity stimuli to assess its ability to discern and compare quantities. To benchmark ViTs against established architectures, we will conduct direct performance comparisons with CNNs, specifically VGG19, and ground our findings by contrasting them with human behavioral data.

Thorough analysis of the operation of the models is critical. Interpreting neural processes at specific layers, a common practice in CNN analysis, becomes more challenging with ViTs. In response, we propose utilizing cumulative attention maps from the ViT, providing insights into the areas of focus by the model during our experiments.

In summary, this research embarks on a novel exploration of Vision Transformers in the context of numerosity discrimination, aiming to unravel whether these models can replicate the cognitive alignment observed in CNNs and, by extension, contribute to our broader comprehension of artificial intelligence's numerical cognition capabilities.

Methods

Models

For our vision transformer, we used a pre-trained base-sized ViT model from Google available on HuggingFace. This model is pre-trained on ImageNet-21k (14 million images, 21,800 classes) and has 86.6 million parameters. We believed that the training objective in this scenario was distinct enough from the numerosity concepts of interest to us, which would show more clearly whether numerosity is truly learnt 'for free' by the architecture.

For our CNN baseline, we use a pre-trained VGG-19 model from the torchvision library. This model is pre-trained on ImageNet-1k (14 million images, 1,000 classes) and contains

143.6 million parameters. Upadhyay (2023) uses the same architecture in their work on CNN numerosity sensitivity, along with several other models including AlexNet and ResNet. However, VGG-19 saw some of the best results overall with respect to R^2 values for the ratio effect, especially in experiment 6 (numerosity in more complex images) where the models had struggled the most in performance. Thus, we chose to use VGG-19 as it would provide some of the most competitive results to compare our ViT against.

Numerosity Data

We use essentially the same procedure for generating the six experiments described in Upadhyay (2023). Each experiment has several images at each numerosity number from 1 through 9 (inclusive). In the first five experiments, the stimuli are a collection of solid black shapes randomly positioned on a white background. Over these five experiments, the area, relative size, and shape type (circle, triangle, square) was altered. For experiment 6, rather than using random items from Google Images as described in the paper, we collected pictures from free numerosity worksheets at the kindergarten level. Examples of input stimuli from these experiments can be seen in Figure 1.

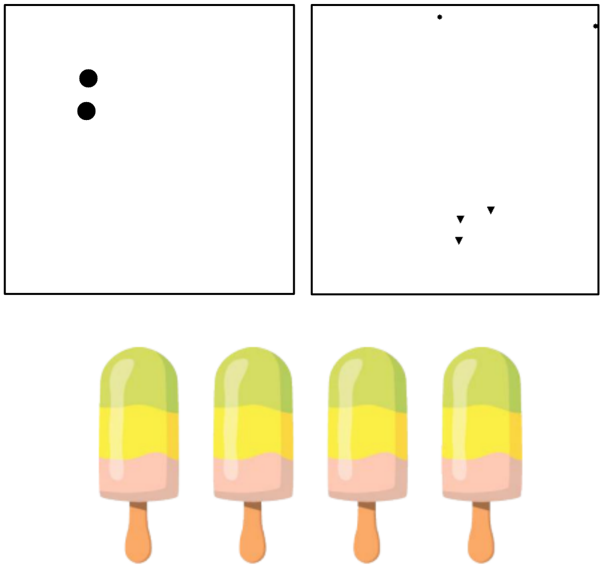


Figure 1: Example stimuli from experiment 2 (top left), experiment 5 (top right), and experiment 6 (bottom), with numerosities of 2, 5, and 4, respectively.

Experiments

Inference, Ratio Effect, And MDS Analysis

We used the hidden layer outputs of our ViT model for the purpose of inference. Specifically, we used the outputs of the last layer (which had dimensionality 768 per token) and used the model's hidden representation of the classification token for similarity/distance representations. We also tried using the hidden representation for all the tokens instead of the classification token; however, the results were not strongly indicative of trends, as seen below in Figure 2. Further, we found that it was more intuitive for the classification token to be used because its embedding attempts to represent the whole picture holistically.

As discussed earlier, the ratio effect informs us about the model's comprehension of the relationship between numerosities: specifically, the understanding of the similarity in these numerosities' representations based on how close they are in ratio. This ratio is calculated simply as max(n1, n2) / min(n1, n2), where n1, n2 are two given numerosities. For the ratio effect plots, we use the metrics.pairwise.cosine_similarity function from sklearn to determine the cosine similarity in the activations. We only present the ratio effect plots for the sake of brevity in this paper.

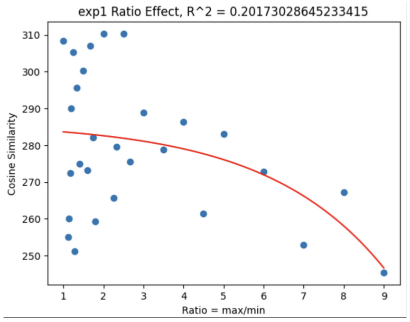


Figure 2: Sample ratio effect graph when using all tokens

We are also interested in looking at the mental number line representations of our ViT for comparison purposes. We estimate this representation by making a 1-dimensional MDS plot based on the activations from each set of experiments. For the MDS plots, we use the spatial.distance.cosine function from the scipy library to get distances in representation for the pairs of numbers. We compare this plot against the ones provided by Upadhyay (2023) for the VGG-19 and human mind mental number lines.

Attention Maps and Grad-CAM Maps

The HuggingFace transformers package allows for the extraction of hidden layer ouptuts and attention layer outputs during inference time via an argument, so no model hooks were required for obtaining these ViT outputs. After collecting the attention weights, we considered several diverse ways to visualize them. First, we attempted simpler methods (e.g. summing across layers and heads, just using weights from specific layers). Further literature review uncovered prior work by Abnar and Zuidema (2020) on quantifying attention flow in transformers. There were repositories that implemented this paper, but we found that the best results were achieved via a simple reimplementation that multiplied the attention weight matrices. Prior to this multiplication process, the weights of a single layer were combined across heads via the max() function; this way, every patch's 'attention' value is the maximum across all heads for a single layer. We finally used the first column of this attention matrix (which corresponds to the classification token) to see which of the patches were most important in determining the classification. This final column was reshaped, scaled up, and then overlaid with the images to get the attention visualizations.

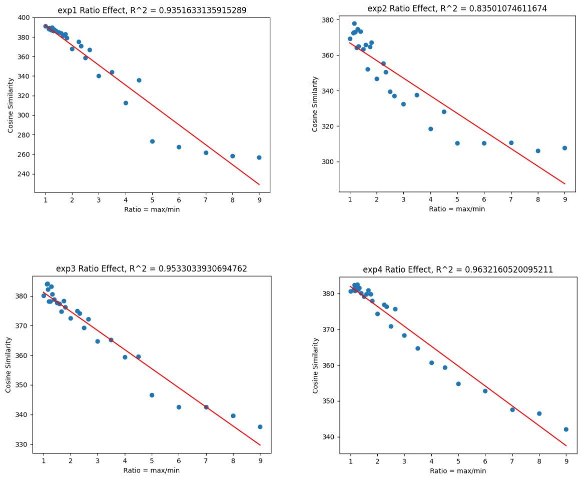
Similarly, we use Grad-CAM (Gradient Class Activation Maps) for analyzing which parts of the image our CNN architectures pay the most attention to. For the VGG-19 model activations, we registered a backward hook to the last convolutional layer of our model (in this case, the 35th layer of the feature block). To obtain the activations, we set the model in evaluation mode and run our model on the input numerosity image. We then run a backward pass and use our hook to obtain the gradients and final convolution layer activations. After weighting the activation channels by their gradients, averaging the channels, and normalizing, we obtain a heatmap which we overlay on top of the input image.

Results

The results for our six experiments are summarized below, with higher level analysis of the models' understanding of numerosity followed by analysis of different numerosities using attention maps and activation maps.

Ratio Effect

The graphs for the ViT ratio effect experiments can be seen in Figure 3; these graphs are comparing the average cosine similarity for each possible ratio. Table 1 shows a comparison between the results for VGG-19 in Upadhyay (2023) and for our ViT with respect to the R^2 value of the best-fit curve.



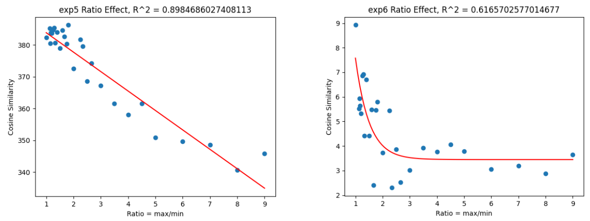


Figure 3: Ratio-effect graphs for the 6 ViT experiments, with best-fit curve. The graphs depicted are in order (left to right, top to bottom) and are titled with the R^2 value.

Table 1: Comparison of R^2 Values For Ratio Effect

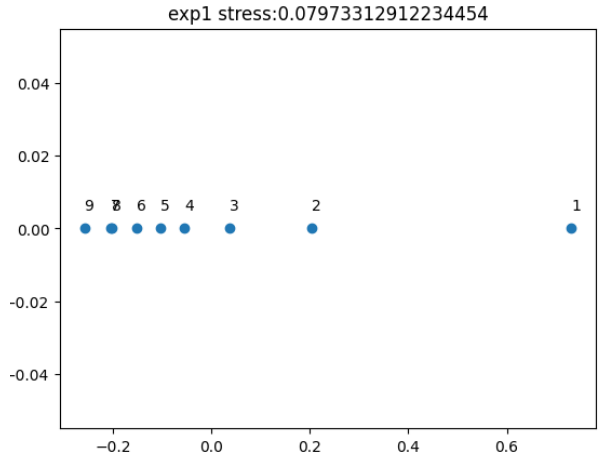
Experiment	VGG-19 R^2	ViT R^2
1	0.9954	0.9352
2	0.9133	0.8350
3	0.9226	0.9533
4	0.7815	0.9632
5	0.9235	0.8985
6	0.4892	0.6166

Figure 3 demonstrates a clearly observable relationship between the cosine similarity and the ratio in numerosities.

As seen from the results above, the R^2 values for the best-fit curve of both models' results are comparable. Each model demonstrates a higher R^2 value in 3 of the 6 experiments. However, the ViT shows a significantly greater R^2 value in experiment 6, which is the most challenging of the experiences. Further analysis of these results is provided in the Discussion section.

Mental Number Line Representations

The mental number line (MNL) representations for the ViT experiment 1 are presented in Figure 4, along with the MNL representations from the VGG-19 model and human mind (provided from Upadhyay 2023).



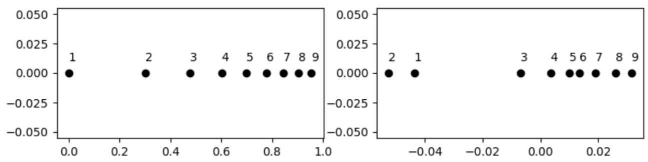


Figure 4: Mental Number Line representations from the ViT (top), human mind (bottom left) and VGG-19 (bottom right). The depicted number lines are for experiment 1.

Note that the direction of the number lines do not matter in this case, rather the relative locations and space between numbers is of importance. The ViT model more clearly mimics the human mind number line than the VGG-19 model, especially at numerosities 1 and 2; the ViT MNL correctly places a numerosity of 2 between 1 and 3.

Attention And Activation Maps

Both the vision transformer and VGG-19 model struggled the most with the input data from experiment 6, however the ViT model did significantly better (26% higher R^2 fit for an exponential function). To further compare their performance on these images, we passed in images from experiment 6 into our models and then extracted the attention maps from the ViT and the class activation maps from the CNN, presented in Figure 5 and Figure 6 respectively.

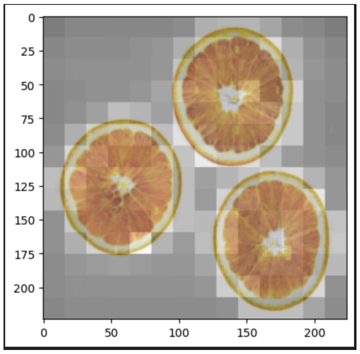


Figure 5: Attention Map from the Vision Transformer. Lighter is more attention.

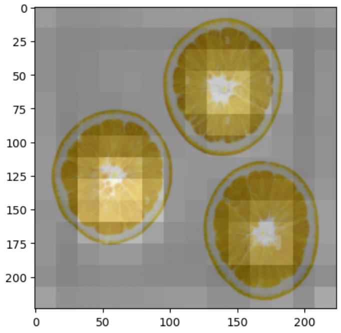


Figure 6: Activation Map from the VGG-19 Model. Lighter is higher activation value.

From visual inspection, it is clear that the attention maps from the vision transformer are focused on each object and in its entirety. Meanwhile, the activations of the VGG-19 for the same image are focused on the center of the object rather than the entire thing. They also appear a lot less consistent per item – the leftmost object in Figure 6 seems to have higher (lighter) activation values than the rightmost objects.

Discussion

In our results, the vision transformer was indeed able to demonstrate numerosity understanding throughout the variety of experiments. The ViT graphs in Figure 3 clearly demonstrate the ratio effect. For experiments 1 through 5, the average of the R^2 values of the best-fit curves is 0.9170, with the lowest R^2 value being 0.8350. This shows that the variation in the ratio explains the variation in cosine similarity quite well. These results show that the vision transformer is indeed picking up on numerosity and is able to visually discriminate between items of different quantities.

The performance was fairly consistent throughout experiments 1 through 5, though the model did struggle more with experiment 6. This is likely explained by the fact that the earlier experiments had a consistent selection of shapes and had high contrast against their backgrounds (always black shapes on white backgrounds); in experiment 6, the objects depicted were significantly more complex, larger (relative to the total image size), and had less contrast with their background.

Further, the results in Table 1 compare the performance of the VGG-19 and ViT models. The performance between both models in discriminating numerosities is comparable, with the ViT results having a higher average of 0.9170 versus the VGG-19 model's average of 0.9073. The VGG-19 model's lowest R^2 value is 0.7815, which is substantially lower than the ViT's lowest value of 0.8350. Additionally, Table 1 shows that the range of the vision transformer R^2 values is much tighter, which, combined with the higher average, shows that it may be more reliable and consistent than the CNN in understanding numerosity concepts without explicit training. This is an important finding because vision transformers do not have the same inductive bias that convolutional neural networks do. Additionally, these results show that the VGG-19 model similarly struggles the most with experiment 6 as well. Figure 7 below shows the ratio-effect graph of VGG-19 for experiment 6, as presented in Upadhyay (2023).

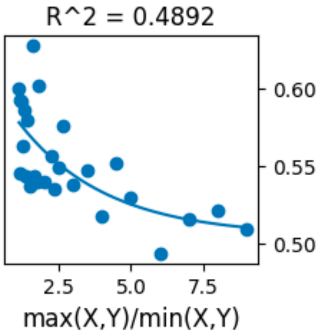


Figure 7: Ratio effect of VGG-19 for experiment 6, where the vertical axis represents the cosine-similarity.

Comparing Figure 7 to the experiment 6 graph in Figure 3, along with their respective R^2 values, we can see that the ViT outperforms VGG-19 by a significant amount. This experiment is more interesting to us since the collection of objects is much more diverse and realistic to common, everyday objects – like the ones that would be present in the training sets of these models. One explanation for why the ViT might outperform here could be because it is trained with ImageNet-21k instead of ImageNet-1k; despite the same amount of training images, the ViT is exposed to a larger set of classes. Another explanation, however, could be that the ViT is learning more generalized representations. This would support our hypothesis that numerosity discrimination is a learned concept, since exhaustive pre-training with objectives is a key part of vision transformers.

Conclusion

Ultimately, the vision transformer does show impressive performance in numerosity comprehension. The results we observe indicate that this model's performance is competitive with Convolutional Neural Network architectures at discriminating numerosities in multiple simpler experiments and even significantly outperforms with experiments involving more realistic objects. The mental number line created by the vision transformer is also more similar to the one produced by the human mind than is the number line produced from the VGG-19. This supports the argument that the concept of numerosity is learned, since despite lacking the CNN's intrinsic inductive bias in working with varying scales and modeling local pixel-level structures, the vision transformer's general representations can still outperform. Lastly, our use of attention maps and activation maps from our models provided some insight into what these models are looking at and how that can explain differences in our results.

There are many promising future directions to take this work. A key area of further research would be to observe the development of this numerosity comprehension in the vision transformer as it trains. For example, one could train the vision transformer from scratch and checkpoint a diverse range of epochs, repeating these numerosity experiments and attention analyses on these model checkpoints. The performance can be compared against prior research on human development of numerosity as well as the results of a CNN architecture trained from scratch.

Further, vision transformers that achieve state of the art performance are notoriously data hungry in training. Since numerosity is a relatively simple cognitive concept, it would also be interesting to analyze how a vision transformer trained on a significantly smaller dataset would fare against the provided results.

References

Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

Halberda, Justin, Michèle MM Mazzocco, and Lisa Feigenson. "Individual differences in non-verbal number acuity correlate with maths achievement." *Nature* 455.7213 (2008): 665-668.

L Kozachkov, KV Kastanenka, D Krotov. Building transformers from neurons and astrocytes. Proceedings of the National Academy of Sciences, 2023

Spelke, Elizabeth S and Kinzler, Katherine D. "Core knowledge." *Developmental science, 10(1):89–96* (2007).

Stoianov, I. and Zorzi, M. "Emergence of a 'visual number sense' in hierarchical generative models." *Nat Neurosci 15, 194–196* (2012). https://doi.org/10.1038/nn.2996.

Testolin, Alberto, et al. "Numerosity discrimination in deep neural networks: Initial competence, developmental refinement and experience statistics." *Developmental science 23.5* (2020): e12940.

Upadhyay, Neha, and Varma, Sashank. "CNN models' sensitivity to numerosity concepts." *The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS'23* (2023).

Wang, Z., & Wu, L. (2023). Theoretical Analysis of Inductive Biases in Deep Convolutional Networks. *arXiv preprint arXiv:2305.08404*.