CNNs Generalize Numerosity Across Naturalistic Stimuli Without Single-Unit Selectivity

Tammo N.H.R. Brandes (tammobrandes@gmx.de)*
Research Master Brain and Cognitive Sciences, Institute for Interdisciplinary Studies. University of Amsterdam

Jisk J. Groot (jiskgroot@gmail.com)*
Research Master Brain and Cognitive Sciences, Institute for Interdisciplinary Studies. University of Amsterdam

Raquel G. Alhama (rgalhama@uva.nl)
Institute for Logic, Language and Computation. University of Amsterdam

Abstract

Previous studies observed that neural network models develop numerosity-selective units when trained to perform object classification, without explicit training on numerosity. However, the emergentist view was challenged by the finding that selectivity disappears with larger sample sizes for model evaluation. Here, we investigate whether this finding was due to the qualitative visual mismatch between training and evaluation data. We present experiments with three types of neural networks, optimized either for object classification, numerosity, or both. Using a novel dataset in which both training and evaluation images include daily-life objects, we analyze layer and single-unit selectivity on a range of conditions, varying the visual properties of our evaluation images. Our results suggest that numerosity classification performance is exclusive to numerosity trained networks. Moreover, we observe a discrepancy between single-unit numerosity selectivity, compared to overall network performance. This suggests that numerosity may be represented through different encoding patterns than previously assumed.

Keywords: cognitive neuroscience; numerosity; vision; computational modeling; neural networks

Introduction

Understanding numbers is an important asset for any biological being. Numerosity estimation—the ability to estimate the number of items in a set-is an evolutionarily ancient skill found across many species (Agrillo, 2015) and even in newborns (Izard, Sann, Spelke, & Streri, 2009). This ability relies on the approximate number system (ANS), which follows Weber's law: sensitivity is ratio-dependent (Feigenson, Dehaene, & Spelke, 2004). Specifically, distinguishing numerosities is easier when they are more different (distance effect) and when they are smaller (size effect) (Gillman & Buckley, 1974). These effects raise questions about how numerosity is constructed in the brain and how these behavioral effects emerge.

Neuroscientists have attempted to identify the neural underpinnings of numerosity estimation by comparing the brain's response to various displays of numerosities. A frequently identified region in these studies is the intraparietal sulcus in which neurons show Gaussian tuning curves to different numerosities (Hubbard, Piazza, Pinel, & Dehaene, 2005). Learning mechanisms have also been studied through computational models. Verguts and Fias (2004) proposed a neural network model where hidden units acted as 'summation units', signaling quantity through graded firing without preferring specific numerosities. More recently, computational efforts to study numerosity have largely shifted to using Convolutional Neural Networks (CNNs, LeCun et al., 1998), which are powerful tools for modeling human vision at the neural level (Güçlü & van Gerven, 2015; Yamins & DiCarlo, 2016). While CNNs are usually trained to classify qualitative information of a visual input, these networks may also be useful to reveal how numerosity estimation is realized in the brain by examining whether numerosity can emerge as a byproduct of visual feature learning. Analyzing tuning curves of network units reveals computational processes and their relation to performance, offering insights into potential shared mechanisms with neural processing. Using this approach, a number of studies observed that numerosity-selective units (that is, network units that respond to specific numbers) can indeed be detected in CNNs trained on object classification in natural images (DeWind, 2019; Nasr, Viswanathan, & Nieder, 2019; Testolin, Zou, & McClelland, 2020); furthermore, these models show similar representation biases compared to humans (Wencheng et al., 2023) and to biological neurons in monkeys (Nasr et al., 2019; Nieder & Merten, 2007).

These findings were challenged when Zhang and Wu (2020) showed that increasing the sample size of the test images decreased the numerosity classification performance of the models. Therefore, numerosity may have been observed in prior research as a result of mere statistical coincidence, rather than being a property that reliably emerged from CNNs trained on object classification over natural images.

The experimental design of the aforementioned study uses natural images as training data and dot-displays for evaluation (where the total number of dots represents the number in the image). Results indicate that deep CNNs only perform well in testing data that is from the same distribution as the training data, except for the numbers 1,2 and 4, which show high generalizability due to several network units showing selective responsiveness to low numbers. Hence a possible explanation for the poor performance of neural networks trained on natural images in numerosity tasks is the mismatch between training and evaluation data. These models were not exposed to dot-displays or similar images during training. Thus, numerosity may still be represented but only detectable in images closer to natural scenes. In other words, the model may still capture numerosity but fail to recognize an image as an

* These authors contributed equally to this work.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

instance of a certain numerosity class due to the inherent difficulty of neural networks to generalize outside the training space (Marcus, 1998; Alhama & Zuidema, 2019).

To address this possible confound, we designed a study where, instead of using dot displays, we created images representing actual everyday objects. The out-of-distribution dataset used new, previously unseen everyday objects. By maintaining consistency between training and testing classes, we expect the model to generalize more effectively.

Using this dataset, we trained CNNs on either object recognition, numerosity, or both (in a multitask setting). We tested the three models on their ability to identify numerosity in the images, where the number was represented by the quantity of objects shown. We performed our evaluation over data sampled under three conditions: within-distribution, out-of-distribution, and using constant object size; in all cases ensuring a large-enough sample size (five times larger than used by Zhang and Wu (2020)). Finally, we explored whether the models showed numerosity-tuned network layers and single units.

Our findings suggest that only networks explicitly trained for numerosity classification show numerosity representations. Moreover, we identified a discrepancy between numerosity selectivity at the level of individual units and the overall network performance. This suggests that numerosity may be encoded through different representational patterns than previously thought.

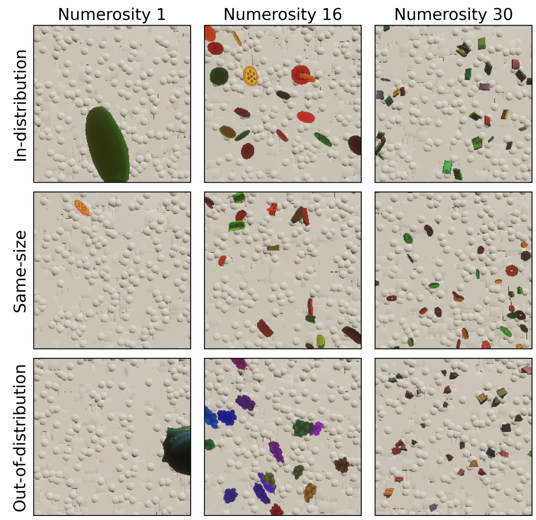

Figure 1: Sample images from our three conditions.

Methods

Network Architecture and Training

We investigated the representation of numerosity in CNNs using three model types: one trained solely on object recognition (ObjNet), one trained on numerosity classification (NumNet), and a multi-task model (MultiTaskNet) trained on both object recognition and numerosity classification.

Based on the study by Nasr et al. (2019), we used an 8-layer convolutional neural network as our backbone with two fully-connected task heads. We used Cross-Entropy Loss for both task heads. MultiTaskNet was trained on the sum of the individual task losses, whereas the single-task networks were trained on their respective task loss only. We used an ADAM optimizer (Kingma & Ba, 2014) with a learning rate of lr = 1e-4 for the MultiTaskNet and ObjNet. For the NumNet, we found that a smaller learning rate of lr = 2.5e-5 achieved better performance. Each network was trained with a batch size of 32 for 20 epochs, repeated four times, yielding a total of 12 networks. For each network, the best-performing epoch based on validation loss was selected for analysis.

Datasets

Three custom synthetic datasets were created using Unity's Perception package (Borkman et al., 2021) to resemble the different dataset conditions of Zhang and Wu (2020)[1]. The in-distribution dataset contained 10 different real-life objects and food items (e.g., a bottle, milk carton, a donut) at the 16 numerosities used by Nasr et al. (2019) with individual objects uniformly scaled according to the numerosity of each image to keep the overall covered area constant. Object color was randomized uniformly for each object in an image using a color mask to avoid color cues. The same-size dataset contained the same objects and the same numerosities but individual objects were not globally scaled according to numerosity. Lastly, the out-of-distribution dataset contained 10 unseen objects, each scaled according to the numerosity of each image similar to the in-distribution dataset. In addition, each individual object instance was randomly scaled, either uniformly (factor 0.8 -1.2 for in-distribution dataset) or per axis (factor 0.6 - 1.4 for same-size and out-of-distribution dataset). All images are of size 224x224x3 and objects were presented against a randomized grey background (see Figure 1 for some sample images).

Decoding Analysis

To investigate the emergence of numerosity and object representations, we evaluated the layer-wise performance of the network. We selected the median performing network from each network type by comparing their lowest achieved validation losses. For each layer, we trained a decoder of the same architecture as the task heads of the networks. The decoders were trained using ADAM with an increasing learning rate for shallower layers proportional to the amount of inputs of each layer to the decoder (starting with lr = 2.5e-5 for 768 channels from layer 8) for 30 epochs, with a patience of 5 epochs. Based on validation performance, the best network was selected for testing. Each network was tested five times in total: once for each dataset, with decoders trained on said dataset to assess maximal decoding performance, and once on each

[1]Code and data supporting this paper are available here: https://github.com/tammobrandes/numerosity-in-anns

of the generalization datasets using the in-distribution-trained decoder network (full-transfer).

Single-Unit Analysis

To investigate whether single units within the networks displayed numerosity-specific responses, we conducted an analysis of unit activation across a range of numerosities. Activation maps were extracted from network units at each network layer. ReLU non-linearity was applied to the activation maps, ensuring that only positive responses were included in the analysis. We evaluated whether units were numerosity selective by applying a similar analysis as used in Nasr (2019). Here, a two-way Analysis of Variance (ANOVA) was performed taking numerosity, dataset and the interaction between numerosity and dataset as factors for prediction unit activation ($\alpha < 0.01$). Units were labeled as numerosity selective if activation was significantly explained by numerosity, and the dataset and the interaction factor were insignificant. To gain insights into the overall numerosity representation within the network, we analyzed the distribution of preferred numerosities across all selective units.

Results

Overall Performance

To examine overall model performance, we computed the macro F1-scores of each network on both tasks and for all datasets (see Figure 2). We used an ANOVA to assess for differences in performance, including only the explicitly trained networks in each individual ANOVA (e.g. for numerosity classification only the MultiTaskNet and the NumNet were included in the analysis).

For object classification, we found a main effect of dataset ($F(2, 18) = 8539.570$, $p < .001$) with Tukey-corrected post-hoc tests revealing significant differences between all dataset combinations (all $p < .001$), with the out-of-distribution dataset yielding the worst performance, followed by the same-size dataset and the in-distribution dataset. In addition, we found a main effect of network ($F(1, 18) = 270.449$, $p <= .001$), representing a significantly lower performance of the MultiTaskNet (M = 0.404) compared to the ObjNet (M = 0.492). Lastly, a significant interaction effect between dataset and network was also found ($F(2, 18) = 153.239$, $p <= .001$), characterized by worse performance on the same-size dataset and in-distribution dataset for the MultiTaskNet compared to the ObjNet (both $p <= .001$). For the out-of-distribution dataset, there was no difference in performance.

For numerosity classification, we also found a main effect of dataset ($F(2, 18) = 370.541$, $p < .001$) characterized by significant drops in performance across all datasets, with the out-of-distribution dataset performing worst followed by the same-size dataset and the in-distribution dataset (all $p < .001$). The main effect of network type was also significant ($F(1, 18) = 11.693$, $p = .003$) with the MultiTaskNet (M = 0.591) outperforming the NumNet (M = 0.572). Lastly, an interaction effect of network type and dataset was found ($F(2, 18) = 6.188$, $p = .009$) due to worse performance of the NumNet (M = 0.639) compared to the MultiTaskNet (M = 0.687) on the in-distribution dataset ($p = .002$).

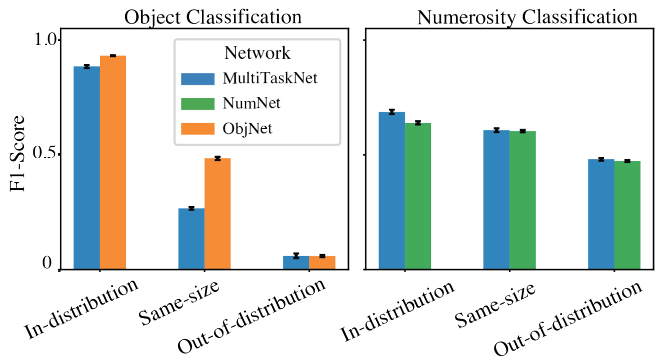

Figure 2: Overall performance (Macro F-1 Score) on classification of object and numerosity, for each type of network.

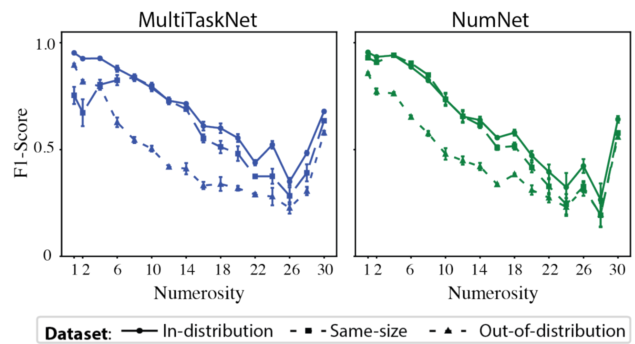

Figure 3: Performance on each numerosity class.

Furthermore, we investigated class-wise performance on the numerosity task in both the NumNet and MultiTaskNet to gain a deeper understanding of the previously identified effects (see Figure 3). The results reveal a gradual decrease in performance from low to higher numerosities, as well as an increase for the two highest numerosities, which might reflect an edge effect due to activation curve competition. A comparison between the datasets revealed that the in-distribution dataset only outperformed the same-size dataset on numerosity 2 ($p = .002$). In contrast, the same-size dataset outperformed the out-of-distribution dataset on most of the intermediate numerosities (6-20, all $p < .005$). Regarding the main effect of network type and its interaction with dataset, class-wise analysis reveals significantly higher F1-scores of the MultiTaskNet compared to the NumNet on the in-distribution data only on numerosities 24 and 28 (all $p <= 0.01$).

Decoding Analysis

To examine how information is represented across network layers, we conducted a decoding analysis using the median-performing model iteration from each network type, as determined by validation loss and decoded both object and numerosity representations. As can be seen in Figure 4, both the

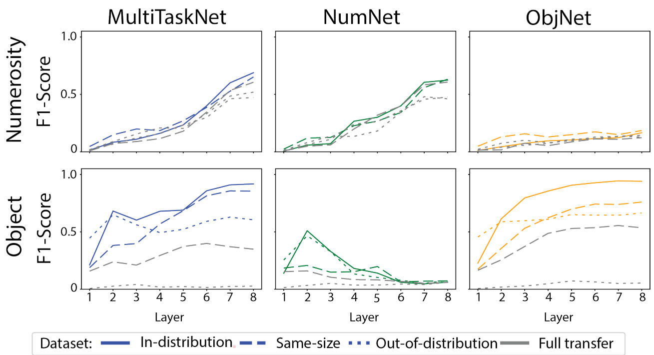

Figure 4: Decoding Analysis: Macro-F1 scores and mean absolute error for each network layer. Line pattern indicates dataset distribution; grey lines depict full transfer.

MultiTaskNet and NumNet show a similar pattern of performance increase in numerosity classification for deeper layers across all datasets. In contrast, the ObjNet showed smaller performance gains in deeper layers compared to the other networks. As revealed in Figure 5, performance increases in both the MultiTaskNet and NumNet were due to improvements in classifying lower numerosities, with performance decreasing as numerosity increased, followed by an increase for the highest numerosities in both the MultiTaskNet and NumNet. The findings also show that the edge effect, previously identified in the overall performance analysis, is also present in the earlier layers of the networks.

Furthermore, we assessed the impact of the datasets on decoding performance across layers and numerosities. The results reveal that the network representations were robust against changes in item size, as seen by the overlapping lines for the in-distribution and same-size datasets across all layers and confirming prior findings from the overall performance analysis. Regarding the performance deficit on the out-of-distribution dataset, we find that this deficit becomes more pronounced in later layers (layer 7 and layer 8) and more so for low to moderate numerosities.

Lastly, we decoded the object representations present in each network layer. In Figure 4 we see a gradual increase in decoding performance across layers for both the MultiTaskNet and the ObjNet. Furthermore, the results show that the low generalization to the same-size and out-of-distribution dataset (as seen by the full-transfer conditions) can be offset by training the decoders on these datasets specifically (as seen by the colored lines). This difference in performance suggests that the networks sufficiently distinguish the objects in the same-size and out-of-distribution dataset but rely on different representations to do so. Lastly, the results show that the NumNet contains object representations in its earlier layers, particularly in layer 2, which are no longer present in later layers. Furthermore, the lower performance of the same-size dataset compared to the out-of-distribution dataset in layer 2 suggests that these representations are largely specific to the numerosity-size relationship in the datasets (i.e. smaller item size with higher numerosities).

Single-Unit Analysis

Of the 37632 network units in the final layer, 71 (0.19%) were identified as numerosity selective in Multitask Net, 19 (0.05%) in NumNet and 147 (0.39%) in ObjNet. As can be seen in Figure 6 (A-C), each network unit exhibited a maximal response to a specific numerosity value (which we refer to as the unit's *preferred* numerosity). The distribution of the preferred numerosities (Figure 6 (D)) differed markedly between the models. MultiTaskNet and ObjNet showed strong biases toward small numbers, with more than 70% units in each network tuned to number 1. ObjNet also exhibited units tuned to numbers 2–6, while MultiTaskNet showed some units tuned to numbers 16-20 and 30. In contrast, the numerosity-selective units in NumNet were only tuned to large numbers 28 and predominantly 30.

Overall, each network exhibited only a small proportion of numerosity-selective units, which did not span the full range of values presented. This suggests that specialized representations of numerosity are not present on the single-unit level.

Discussion

Is object recognition a sufficient guiding principle for the emergence of numerosity? Zhang and Wu (2020) observed that the percentage of numerosity-selective units in an object-detection CNN, detected using the two-way ANOVA method, decreases significantly as the sample size increases. Specifically, they noted that the 9.6% numerosity selectivity reported by Nasr et al. (2019) dropped to approximately 2% when the sample size was increased to 20 test images per combination, and eventually reached 0% with 80 test images. To address this limitation, we expanded the evaluation to a much larger and more robust dataset: 16,000 test images, comprising 1,000 variations across 16 numerosity levels. Crucially, we maintained consistency in visual characteristics between training and test sets by using naturalistic images of everyday objects throughout, thereby avoiding the distribution mismatch observed in Zhang and Wu (2020) where training used natural images and testing used dot displays. In contrast to dot-displays used by prior studies, our dataset required the networks to build object-independent representations of numerosity. Despite this more ecologically valid design and increased statistical power ObjNet exhibited almost no numerosity-selective units, with only 0.39% passing the selectivity criterion, and these were restricted to the lowest numerosities (1–4). This minimal selectivity was mirrored in the network's poor performance on numerosity classification tasks across all layers. These findings provide strong evidence that training on object recognition alone is insufficient for the emergence of numerosity representations, even when using large, naturalistic datasets and a consistent training-test distribution.

Our single-unit analysis aligns with the core finding from Zhang and Wu (2020), showing that all networks, irrespec-

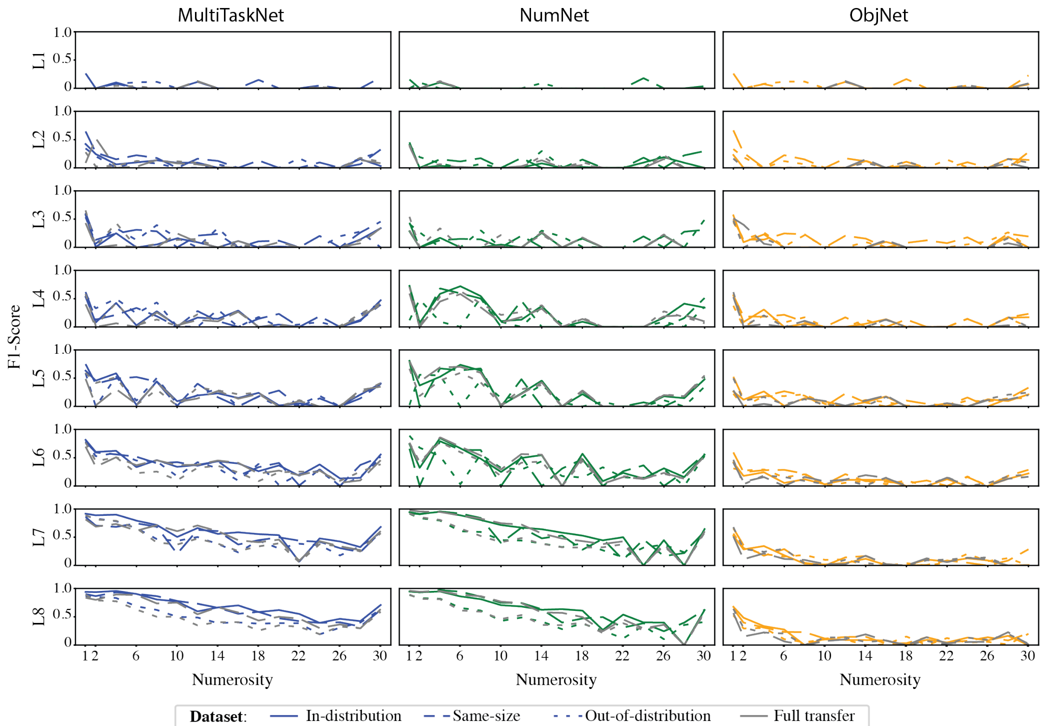

Figure 5: Decoding Analysis: Micro F1-Scores per Layer

tive of the training objective, contain very few numerosity-selective units when evaluated on a large dataset. However, in contrast to their results, we observe that both MultiTaskNet and NumNet achieve high classification accuracy and generalize effectively to out-of-distribution stimuli. This discrepancy yields two important insights. First, it demonstrates that convolutional networks are indeed capable of learning representations of numerosity that generalize to out-of-distribution data. This capacity for generalization may be enhanced by the use of naturalistic object stimuli, which introduce greater variability and ecological validity than the uniform dot displays employed in earlier studies. This is in line with Chapalain, Thirion, and Eger (2024), who argue that these previous networks were likely sensitive to low-level features of the dot-displays rather than numerosity. When tested on realistic input images, numerosity-selective units, as identified by dot-displays, failed to show above-chance decoding performance. This highlights the importance of training data diversity in promoting numerosity representations in CNNs. Second, and perhaps more strikingly, our results show that strong classification performance does not require a large number of numerosity-selective units. This raises a central question: How can CNNs achieve reliable numerosity inference despite the apparent absence of a large number of units explicitly tuned to specific numerosities?

A plausible explanation is that numerosity is not encoded by isolated single-units, but rather emerges from distributed patterns of activity across the network. In this view, classification performance is supported not by specialized numerosity-selective units, but by population-level representations of numerosity. A similar finding has been reported in Chapalain et al. (2024), where excluding numerosity-selective units led to better numerosity decoding in an object-trained network. Notably, this network was able to form numerosity representations, in contrast to our object-trained network, likely because it had to discriminate between 6 broadly spaced numerosities, in comparison to our network which had to discriminate between 16 closely spaced numerosities. Thus, the apparent lack of numerosity representations in the ObjNet compared to the network used by Chapalain et al. (2024) could be a result of task difficulty rather than a genuine lack of numerosity representations.

Several other studies have proven that numerosity could be encoded through population encoding, without the need for numerosity-selective single-units. For example, it has been demonstrated that on-center/off-surround dynamics–a pattern of activation where the center of a receptive field excites and the surround inhibits, or vice versa, enhancing contrast and edge detection–could approximate numerosity purely through the collective activity of its units, relying on general visual processing principles rather than specific single-unit tuning (Sengupta, Surampudi, & Melcher, 2014; Stoianov & Zorzi, 2012; Park & Huber, 2022).

Our networks might rely on a similar mechanism to represent numerosity. It should be noted, however, that these previous studies used simple dot-displays, and this dynamic might be a byproduct of this simplistic notion of numerosity. Future work should explore whether networks exposed to more

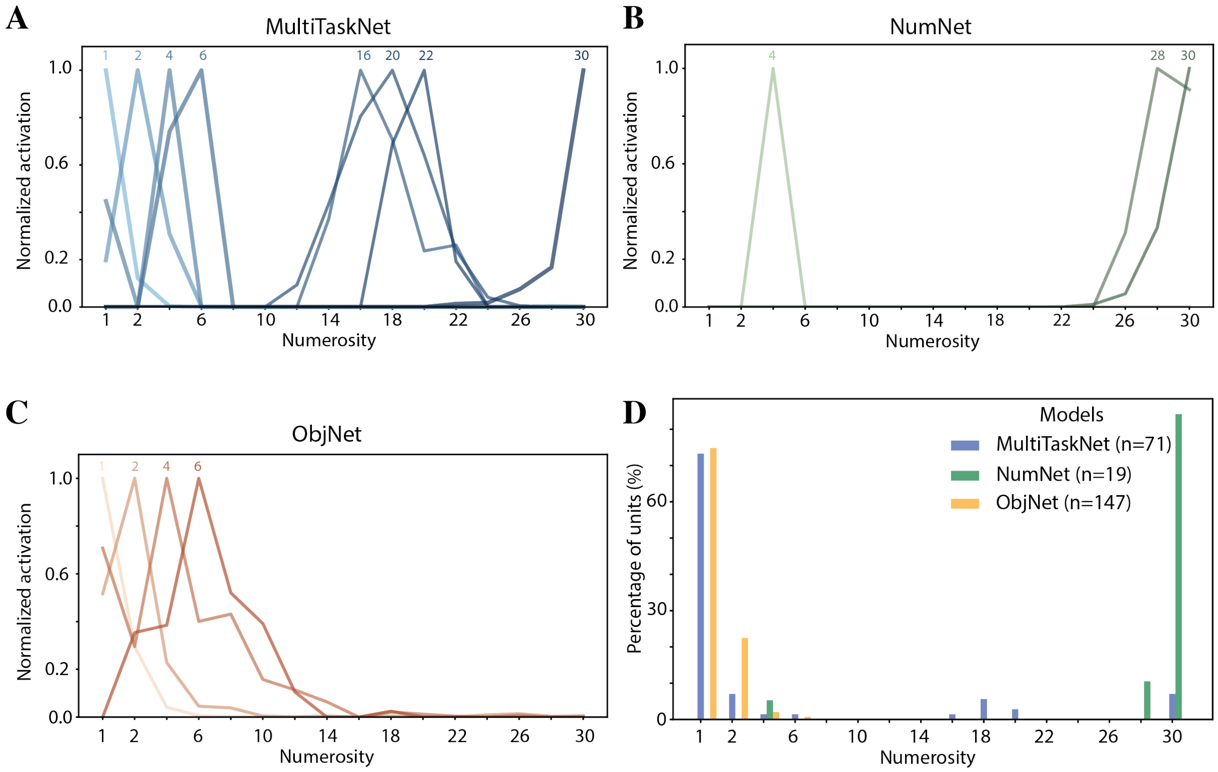

Figure 6: Single-Unit Analysis. Numerosity tuning curves and distributions of numerosity-selective units across three network models. (A-C) Tuning curves of numerosity-selective units color-coded by their preferred numerosity. (A) MultiTaskNet, (B) NumNet, and (C) ObjNet show normalized activation for all numerosities. (D) Histogram of the preferred numerosity distribution across all numerosity-selective units in MultiTaskNet (blue), NumNet (green), and ObjNet (orange). The x-axis represents numerosity, and the y-axis shows the percentage of units tuned to each numerosity.

naturalistic displays exhibit similar dynamics, for instance by using activation maximization to visualize preferred kernel inputs (Yosinski, Clune, Nguyen, Fuchs, & Lipson, 2015).

Taken together, these findings highlight the limitations of current deep learning models (CNNs more specifically) in capturing the abstract nature of numerosity. While numerosity representations can emerge within these networks, they do not necessarily take the form of clearly defined, numerosity-selective units. This aligns with previous concerns that the presence of such units may have been overestimated due to methodological choices and small sample sizes (Zhang & Wu, 2020). Our results show that networks can achieve high accuracy in numerosity classification with very few, if any, dedicated number-selective units, suggesting that population-level encoding mechanisms are likely at play. This could reflect principles similar to those found in biological vision, such as center-surround processing in early visual areas (Hubel & Wiesel, 1962). Using these mechanism numerosity is inferred not from isolated neurons but from the spatial pattern of activity across a network (Sengupta, Surampudi, & Melcher, 2014; Stoianov & Zorzi, 2012). These insights point to the importance of examining distributed representations and their dynamics when evaluating whether and how abstract cognitive capacities emerge in artificial systems.

Conclusion

This study investigated the role of input data and learning objectives in the emergence of numerosity representations within CNNs. By comparing models trained on object recognition, numerosity classification, and a multi-task combination of both, we explored how numerosity sensitivity develops and whether it generalizes to out-of-distribution data.

Our findings indicate that high performance in numerosity classification is limited to networks explicitly trained for that task. Importantly, none of the models, regardless of training objective, contained a substantial population of numerosity-selective units when evaluated using a large, naturalistic test set. This suggests that reliable numerosity estimation does not depend on single-unit numerosity-tuning. Instead, our results point toward the use of distributed population codes as the primary mechanism through which CNNs infer numerosity. This challenges the emphasis on single-unit selectivity as the primary indicator of numerosity representation and supports the view that numerosity can be encoded through coordinated activity across multiple units.

References

Agrillo, C. (2015). Numerical and arithmetic abilities in non-primate species. In R. C. Kadosh & A. Dowker (Eds.), *The oxford handbook of numerical cognition* (pp. 214–236). Oxford University Press. doi: 10.1093/oxfordhb/9780199642342.013.002

Alhama, R. G., & Zuidema, W. (2019, May 28). A review of computational models of basic rule learning: The neural-symbolic debate and beyond. *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-019-01602-z

Borkman, S., Crespi, A., Dhakad, S., Ganguly, S., Hogins, J., Jhang, Y.-C., ... Yadav, N. (2021). *Unity perception: Generate synthetic data for computer vision.* Retrieved from `https://arxiv.org/abs/2107.04259`

Chapalain, T., Thirion, B., & Eger, E. (2024). Trained deep neural network models of the ventral visual pathway encode numerosity with robustness to object and scene identity. *bioRxiv*. doi: 10.1101/2024.09.05.611433

DeWind, N. K. (2019, 4). The number sense is an emergent property of a deep convolutional neural network trained for object recognition. *bioRxiv*. doi: 10.1101/609347

Dobs, K., Martinez, J., Kell, A. J. E., & Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. *Science Advances*. doi: 10.1126/sciadv.abl8913

Feigenson, L., Dehaene, S., & Spelke, E. (2004, 7). Core systems of number. *Trends in Cognitive Sciences*, *8*, 307-314. doi: 10.1016/j.tics.2004.05.002

Gillman, P. B., & Buckley, C. B. (1974). Comparisons of digits and dot patterns. *Journal of Experimental Psychology*, *103*, 1131-1136.

Güçlü, U., & van Gerven, M. A. (2015, 7). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*, 10005-10014. doi: 10.1523/JNEUROSCI.5023-14.2015

Hubbard, E. M., Piazza, M., Pinel, P., & Dehaene, S. (2005, 6). Interactions between number and space in parietal cortex. *Nature Reviews Neuroscience*, *6*, 435-448. doi: 10.1038/NRN1684

Hubel, D. H., & Wiesel, T. N. (1962, 1). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*, 106-154. doi: 10.1113/JPHYSIOL.1962.SP006837

Izard, V. R., Sann, C., Spelke, E. S., & Streri, A. (2009, 6). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 10382-10385. doi: 10.1073/PNAS.0812142106

Kingma, D. P., & Ba, J. L. (2014). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. Retrieved from `https://arxiv.org/abs/1412.6980v9`

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*, 2278–2324. doi: 10.1109/5.726791

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, *37*(3), 243–282. doi: https://doi.org/10.1006/cogp.1998.0694

Nasr, K., Viswanathan, P., & Nieder, A. (2019). Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Science Advances*, *5*. doi: 10.1126/SCIADV.AAV7903

Nieder, A., & Merten, K. (2007). A labeled-line code for small and large numerosities in the monkey prefrontal cortex. *Journal of Neuroscience*, *27*(22), 5986–5993.

Park, J., & Huber, D. E. (2022). A visual sense of number emerges from divisive normalization in a simple center-surround convolutional network. *eLife*, *11*. doi: 10.7554/eLife.80990

Sengupta, R., Surampudi, B. R., & Melcher, D. (2014, 9). A visual sense of number emerges from the dynamics of a recurrent on-center off-surround neural network. *Brain Research*, *1582*, 114-124. doi: 10.1016/j.brainres.2014.03.014

Stoianov, I., & Zorzi, M. (2012, 2). Emergence of a 'visual number sense' in hierarchical generative models. *Nature Neuroscience*, *15*, 194-196. doi: 10.1038/nn.2996

Testolin, A., Zou, W. Y., & McClelland, J. L. (2020, 9). Numerosity discrimination in deep neural networks: Initial competence, developmental refinement and experience statistics. *Developmental science*, *23*. Retrieved from `https://pubmed.ncbi.nlm.nih.gov/31977137/` doi: 10.1111/DESC.12940

Verguts, T., & Fias, W. (2004, 11). Representation of number in animals and humans: A neural model. *Journal of Cognitive Neuroscience*, *16*, 1493-1504. doi: 10.1162/0898929042568497

Wencheng, W., Ge, Y., Zuo, Z., Chen, L., Qin, X., & Zuxiang, L. (2023, 8). Visual number sense for real-world scenes shared by deep neural networks and humans. *Heliyon*, *9*, e18517. doi: 10.1016/J.HELIYON.2023.E18517

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *bioRxiv*. doi: https://arxiv.org/abs/1506.06579

Zhang, X., & Wu, X. (2020). On numerosity of deep neural networks. *Advances in Neural Information Processing Systems*. Retrieved from `https://arxiv.org/abs/2011.08674v1`