Evaluating Visual Number Discrimination in Deep Neural Networks

Ivana Kajić (kivana@deepmind.com)
DeepMind, Montréal, QC, Canada

Aida Nematzadeh (nematzadeh@deepmind.com)
DeepMind, London, United Kingdom

Abstract

The ability to discriminate between large and small quantities is a core aspect of basic numerical competence in both humans and animals. In this work, we examine the extent to which the state-of-the-art neural networks designed for vision exhibit this basic ability. Motivated by studies in animal and infant numerical cognition, we use the numerical bisection procedure to test number discrimination in different families of neural architectures. Our results suggest that vision-specific inductive biases are helpful in numerosity discrimination, as models with such biases have lowest test errors on the task, and often have psychometric curves that qualitatively resemble those of humans and animals performing the task. However, even the strongest models, as measured on standard metrics of performance, fail to discriminate quantities in transfer experiments with differing training and testing conditions, indicating that such inductive biases might not be sufficient.

Basic Numerical Competence

The ability to represent abstract numbers and compare numerical quantities is a basic numerical competence observed in both animals and humans (Dehaene, Dehaene-Lambertz, & Cohen, 1998). It helps animals in foraging, navigation, hunting, and reproduction (Nieder, 2020), and is also correlated with the later mathematical ability in prelinguistic infants (Gilmore, McCarthy, & Spelke, 2007; Halberda, Mazzocco, & Feigenson, 2008). While such a skill is shared across species and is independent of explicit feedback or formal education (Dehaene, 1997; Gallistel & Gelman, 1992), the degree to which more advanced numerical skills, such as counting and symbolic representation of number, are present across species remains a debated topic (O'Shaughnessy, Gibson, & Piantadosi, 2021; Revkin, Piazza, Izard, Cohen, & Dehaene, 2008; Anobile, Cicchini, & Burr, 2016; Gallistel & Gelman, 1992).

To investigate number representation and processing, different neural networks have been used as cognitive models of various numerical skills such as magnitude comparison (Verguts & Fias, 2004; Dehaene & Changeux, 1993; Zorzi & Butterworth, 1999), subitizing (Peterson & Simon, 2000) and counting (Rodriguez, Wiles, & Elman, 1999; Fang, Zhou, Chen, & McClelland, 2018). Neural networks are able to encode exact magnitudes (Creatore, Sabathiel, & Solstad, 2021) and develop basic numerical abilities such as numerosity comparison (Testolin, Dolfi, Rochus, & Zorzi, 2020).

While such networks have been used successfully to explain different phenomena in numerical cognition, their architecture is often designed for a task targeting specific cognitive function. In contrast to such specialized networks, in recent years we have witnessed a radical improvement in both the performance, and the quality of representations learned by deep neural networks that are trained end-to-end across vision (Simonyan & Zisserman, 2014; He, Zhang, Ren, & Sun, 2016), language (Vaswani et al., 2017; Devlin, Chang, Lee, & Toutanova, 2018; Brown et al., 2020), and multimodal (Lu, Batra, Parikh, & Lee, 2019; Radford et al., 2021; Alayrac et al., 2022) domains.

Here, we investigate whether state-of-the-art models designed for visual processing, also referred to as vision encoders, can exhibit basic numerical competence as observed in humans and animals. Specifically, we evaluate *number discrimination* in vision encoders, defined as the ability to make broad relative numerical judgements such as many versus few, which is imprecise and not as advanced as counting, but within the normal ability of many animals (Davis & Memmott, 1982). We draw inspiration from studies in animal and child cognition and use a simple discrimination paradigm known as the *bisection task* to examine if recent vision encoders can learn to discriminate stimuli on the basis of number.

We consider three vision encoders with varying degrees of explicit inductive biases: ResNet (He et al., 2016), ViT (Dosovitskiy et al., 2020), and Swin (Liu et al., 2021), as well a simple, comparatively small, multi-layer perception (MLP) not designed for vision tasks as a baseline. Across all conditions, Swin and ResNet with image-specific inductive biases are the most successful models in number discrimination; moreover, Swin matches the empirical data from humans and animals in more conditions than ResNet suggesting that its additional hierarchical bias results in a better abstract number representation. Even the strongest models, however, often fail in conditions that test for the transfer of numerical skill to a new condition; for example, when models are trained on a stimulus with solid shapes but tested on a stimulus where shapes are not filled. Although models fail in such transfer conditions, we find that they do learn structured number representations, forming clusters that are ordered based on the number identity. This suggests that, unlike humans and animals whose numerical skills generalize across different ecological contexts, vision encoders might require additional modeling innovations or a greater quantity and va-

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

riety of data to use their learned knowledge in new situations.

The Numerical Bisection Task

The numerical bisection task is used to assess perception of numerical quantitites in both animals and humans. First, a participant is trained to discriminate small and large sample numerosities by associating them with different responses (labels such as *few* and *many*). For example, Emmerton, Lohmann, and Niemann (1997) train pigeons to respond to images with 1 or 2 shapes by pecking to the left (corresponding to *few*), and to the right for images with 6 or 7 shapes (corresponding to *many*). In Almeida, Arantes, and Machado (2007), children learn to pick a green cup for 2 drumbeats, or a blue cup for 8 drumbeats in one experiment, and raise a red glove on their left hand after 2 drumbeats and a yellow glove on their right hand after 8 drumbeats in another experiment. The numerosities used for training (*i.e.,* 1, 2, or 8) are often referred to as *anchor numerosities*.

Then, to probe number discrimination, participants are subsequently tested on intermediate numbers that are *not* seen during training (*e.g.,* 3 in the previous experiment). A participant is more likely to select the response associated with the larger anchor value (*e.g., many*), resulting in an s-shaped psychometric curve. Such s-shaped psychometric curves have been used to characterize basic numerical competence in rats (Meck & Church, 1983), pigeons (Honig & Stewart, 1989; Emmerton et al., 1997), rhesus macaques (Jordan & Brannon, 2006), as well as children and adults (Droit-Volet, Clément, & Fayol, 2003; Almeida et al., 2007; Jordan & Brannon, 2006). Qualitatively, psychometric curves documented in the literature have the following characteristics: (1) the initial segment with smaller numerosities is mostly labeled with *few*, (2) intermediate segment with a gradually increasing slope reflecting an increase in *many* responses, (3) final segment with the largest numerosities mostly labeled with *many*. Although these properties characterize the majority of psychometric responses documented in the literature, between- and within-subject variability has been observed depending on the task and numerosity ranges (Almeida et al., 2007).

Experimental Stimuli

We automatically generate images with black background and white circles varying the number of circles from 1 to 7. Similar to Emmerton et al. (1997), we use images with 1, 2, 6, or 7 circles as anchor numerosities for training. When designing stimuli, previous work has identified and controlled for potential perceptual confounds such as the size of the constituent elements (*i.e.,*, circles in our case), total white area, or total perimeter (Honig & Stewart, 1989; Testolin et al., 2020; Emmerton et al., 1997); processing these nonnumerical features—which may be a confound in the observed numerical discrimination behavior—can develop independently of number processing, as has indeed been observed in children's developmental trajectory (Odic, 2018). To control for such potential confounds, we generate six different stimulus categories shown in Figure 1:

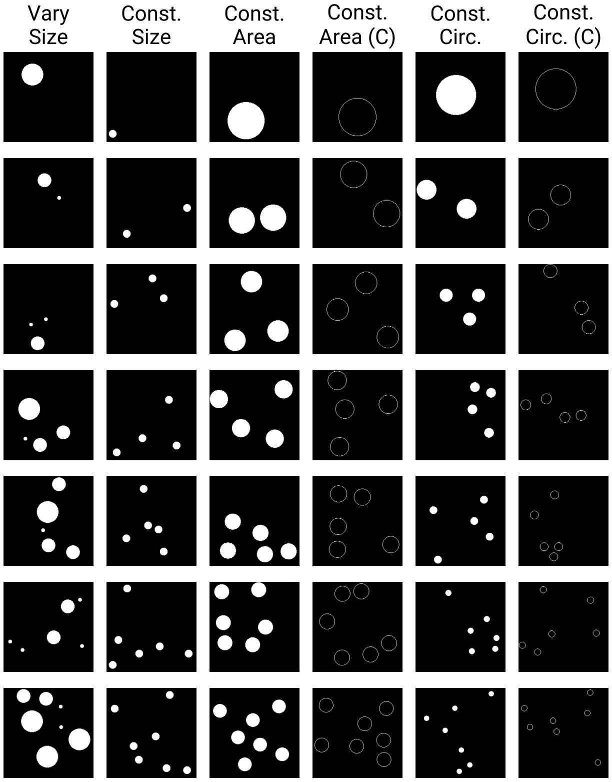

Figure 1: Sample visual stimuli used in the numerical bisection task. Rows are different numerosities and columns different stimuli types. "(C)" denotes "contours", as opposed to shapes with solid white background.

1. *Vary Size*. This is our most general setting, where for each image, we draw circles with radii drawn randomly from a set of 3 values ($r = \{10, 35, 55\}$).

2. *Constant Size*. We control for the size of circles—all circles have the same radius ($r = 20$); this enables us to examine whether models can discriminate numbers better when circles are identical compared to varied in size.

3. *Constant Area*. In the previous condition (*Constant Size*), the white area (covered by circles) increases as the number of circles increases. We control for this potential confound by fixing the total white area to be constant across stimuli. This results in smaller circles in images that depicts larger numbers.

4. *Constant Area (contour)*. We also examine if solid shape background has an impact on models' behavior; we consider a condition the same as *Constant Area*, but using contours instead of shapes with white background.

5. *Constant Circumference*. While the total area is controlled for in the *Constant Size* condition, the total circumference of circles increases with numerosity. Here, we control for the total circumference by keeping it constant across stimuli.

6. *Constant Circumference (contour)*. It is the same as *Constant Circumference*, but using contours instead of shapes with white background.

We generate the stimuli on-the-fly for both the training and testing and store them in-memory to be used during training and testing, with 100 images generated for each numerosity category of one stimulus type, resulting in overall 400 images for training, and 1,100 images for testing for each stimulus type. The images are of dimensionality expected by models, *i.e.,* $224 \times 224 \times 3$.

Experimental Setup

In this section, we examine a few recent families of deep neural networks designed for computer vision (henceforth, vision encoders); all these models have achieved impressive results on computer vision tasks (such as image classification), but differ with respect to inductive biases that their architecture encode. We first describe these models briefly, and then discuss the details of our experimental setup.

Models. We consider three types of vision encoders: ResNet (He et al., 2016), ViT (Dosovitskiy et al., 2020), and Swin (Liu et al., 2021). The ResNet model includes a stack of convolutional neural network (CNN) blocks that process images using convolution kernels. These kernels introduce an explicit *locality* bias—pixels (or features depending on the layer) that are close spatially are combined; as a result, a model with CNN blocks typically learns to encode low-level features (such as edges) in its first layers, and more high-level ones (such as parts) in its last layers.

Both ViT and Swin use Transformer blocks (Vaswani et al., 2017) consisting of feed-forward layers and a *self-attention* mechanism; self-attention introduces a weaker and less explicit *locality* bias (compared to CNNs) as a model can learn to group neighboring image patches.[1] Swin builds on ViT and introduces an explicit *hierarchical* bias by modifying how self-attention is applied across different layers; more specifically, local image patches are merged at at various stages as the depth of the model increases, resulting in a hierarchical representation.

We use specific variants of ResNet, ViT, and Swin encoders: the ResNet-50 variant with 25.6M parameters (He et al., 2016), ViT-B (Dosovitskiy et al., 2020) with 86M parameters, and "tiny" Swin, Swin-T (Liu et al., 2021), with 29M parameters. We picked the smallest ViT and Swin variants, and a ResNet model that has a similar number of parameters to Swin.

Finally, as a simple baseline, we consider a generic feed-forward multi-layer perceptron (MLP) that does not include any inductive biases such as convolutions or attention which are known to be helpful for processing of real-world images. We use an MLP consisting of 2 hidden layers with 256 units each, separated by ReLU non-linearities, and a final linear layer with 2 units. With 0.13M parameters and no "bells and whistles", this makes it a substantially smaller, yet less computationally inexpensive baseline model.

Training. For each stimulus type (*e.g., Constant Area*), we train ResNet, ViT, Swin and MLP models on data generated for that stimulus, *i.e.,* images and their labels (*few* and *many*). More specifically, we add a classification head to these models, to predict the label *few* for images with 1 and 2 circles, and *many* for images with 6 and 7 circles, where labels are encoded as one-hot vectors. All models are trained with a cross-entropy loss and L2 regularization. To get an estimate of variability in model responses for each stimulus category, we train 10 networks by choosing a different seed that randomly initializes network weights.

We perform a hyper-parameter search on the batch size, number of steps, learning rate, and optimizer type to find combinations where training loss has converged on the validation set, and where a network is achieving close to 100% accuracy on the training set. Accuracy is defined as a percentage of correctly classified labels.[2]

Testing. We test the models on new images of anchor numerosities (*i.e., not* seen during training), as well as images of novel interpolated numerosities: 3, 4 and 5. We use 100 images for each numerosity category and each stimulus type.

Experimental Results

In Experiment 1, we investigate models' behavior when trained and tested on the same stimulus type. In Experiment 2, we investigate transfer of the number discrimination skill by testing models on images from a stimulus category that is not used in training (*i.e.,* train on *Constant Size*, and test on *Constant Area*).

Experiment 1: Number Discrimination

In this experiment, we test number discrimination using images from the same stimulus category that is used during training (*e.g.,* train on *Constant Size*, and test on *Constant Size*). We evaluate models based on the accuracy of the stimulus test set, and the quantitative and qualitative characteristics of psychometric curves in relation to the empirical data.

Performance on seen numbers. We first examine the performance of the four architectures when tested on novel images of the anchor numerosities (seen during training): 1, 2, 6, 7. An error occurs when an image with a small numerosity

[1] Self-attention is designed for sequential data such as language; thus, it is less suitable for modelling the two-dimensional spatial relations among image patches.

[2] We find that the batch size of 16, and 5,000 steps worked well for all models, although losses in some models (*e.g.,* Swin and ResNet) converged much faster. We use the Adam optimizer (Kingma & Ba, 2014) for MLP, ViT, and Swin models, with learning rates of 1e-04, 5e-04, and 5e-05, respectively. We use the SGD optimizer with a learning rate of 1e-2 for ResNet. The models are trained using a NVidia Tesla V100 GPU.

	Few (1, 2)				Many (6, 7)				Total Error (Few+Many)			
	ResNet	ViT	Swin	MLP	ResNet	ViT	Swin	MLP	ResNet	ViT	Swin	MLP
Vary Size	3.0	11.8	0.2	**17.4**	3.6	**11.9**	0.6	7.8	3.3	11.8	0.4	**12.6**
Const. Size	0.0	**35.5**	0.0	32.1	0.0	0.3	0.1	**5.0**	0.0	17.9	0.0	**18.6**
Const. Area	1.1	**33.0**	0.0	7.5	0.8	**7.2**	0.1	2.8	0.9	**20.1**	0.1	5.1
Const. Area (C)	0.4	8.0	0.0	**63.2**	1.5	**12.8**	0.3	10.1	0.9	10.4	0.1	**36.6**
Const. Circ.	0.6	0.4	0.0	3.5	0.0	3.5	0.0	**51.8**	0.3	1.9	0.0	**27.6**
Const. Circ. (C)	8.7	**40.8**	0.7	31.1	9.9	35.0	12.8	**44.6**	9.3	**37.9**	6.8	**37.9**

Table 1: Error rates (%) in classifying anchor numerosities as either “few” or “many” on respective test sets. Highest error rates for each stimulus type and each anchor numerosity are highlighted. Highest total error rates across stimuli for each model are underlined.

(*i.e.,* 1 or 2 circles) is classified as *many*, or when an image with a large numerosity (*i.e.,* 6 or 7) is classified as *few*. Average error rates for each network and each stimulus type are shown in Table 1. Overall, we observe that ResNet and Swin, with image-specific inductive biases, have smallest mean error rates of less than 1% in 6/12 and 11/12 conditions, respectively. ViT has mean error rates that are in some cases comparable to or even exceed errors of the MLP baseline. When averaged across all 4 numerosities, we find that highest error rates are consistently observed with the *Constant Circumference (contour)* stimulus category (See Table 1, column “Total error”); suggesting that this combination of visual features represented the most challenging dataset for number abstraction. Meanwhile, no such consistent pattern exists for datasets resulting in smallest errors—the smallest error for ResNet is observed with *Constant Size*, for ViT and Swin with *Constant Circumference*, and for the MLP with *Constant Area*. This observation is not surprising given that these models encode different inductive biases.

Performance on new numbers. Next, we examine how different models perform on numbers not seen at the training time (*i.e.,* 3, 4, and 5). We plot the psychometric curves for selected stimuli, showing percentages of *many* responses across numerosities for models trained on that stimuli in Figure 2. We selected these stimulus categories as representative of the easiest (*Constant Size*, *Constant Circumference*) and hardest (*Constant Circumference (contour)*) conditions based on the average error rates in Table 1. Different from Table 1, each value on the y-axis represents a proportion of *many* responses for a certain numerosity (x-axis).

Overall, some curves in Fig. 2 exhibit characteristics of typical psychometric functions as discussed in Sec. Experimental Setup—specifically, for small numerosities 1, 2, and sometimes 3, we observe a slowly accelerating initial segments, followed by gradual increase with intermediate numbers, and a slowly decelerating final segment for larger numerosities (6, 7). Examples of such curve profiles are Swin and ResNet responses to *Constant Size*, and *Constant Circumference* stimulus categories. However, there are also curves that have atypical flat shapes indicative of failure to learn this task, *i.e.,* those not found in the literature. Out of all 24 curves we analyzed, 4 in total exhibit such a shape with 3 shown in Fig. 2 (*i.e.,* MLP with *Constant Circumference* and MLP with *Constant Circumference (contour)*, ViT with *Constant Circumference (contour)*). Such curves either do not start at, or do not end at expected values indicating lack of sensitivity to number categories, some of which is also evident based on large error values in Table 1.

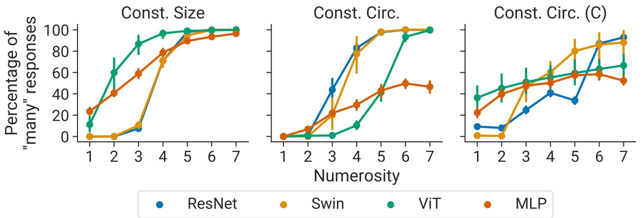

Figure 2: Psychometric functions for each model trained and tested on one stimulus type on the numerical bisection task. Vertical bars are 95% bootstrapped CIs.

Experiment 2: Transfer to Novel Stimuli

The previous experiment demonstrates that Swin and ResNet architectures to a great degree appear to be able to differentiate number of items in an image when trained and tested on the same stimulus type (*e.g.,* constant total area or circumference). To understand whether our models have indeed developed a notion of a number category as opposed to learning a given stimuli, we draw a parallel with research in animal cognition and examine if the models “*base their behaviour on the numerosity of a set, independent of its other attributes*” (Gallistel & Gelman, 1992). In other words, if models learn an abstract representation of a number category, we would expect this representation to be agnostic to perceptual features of the stimulus. To test this, we examine models in a cross-stimulus transfer setting: we train a model on one set of stimuli, but test it on other types of stimuli (*i.e.,* train on *Constant Size*, test on *Constant Circumference*). The test stimuli are *out of distibution (OOD)* with respect to the model’s training distribution. Compared to the *in distribution* setting where training and test are drawn from the same distribution, the OOD setting is known to be challenging for neural networks (Geirhos et al., 2020).[3]

[3] The degree to which a stimulus category is OOD depends on the target distribution, as some stimulus categories are correlated across different dimensions; *e.g.,*, total cumulative white area increases on

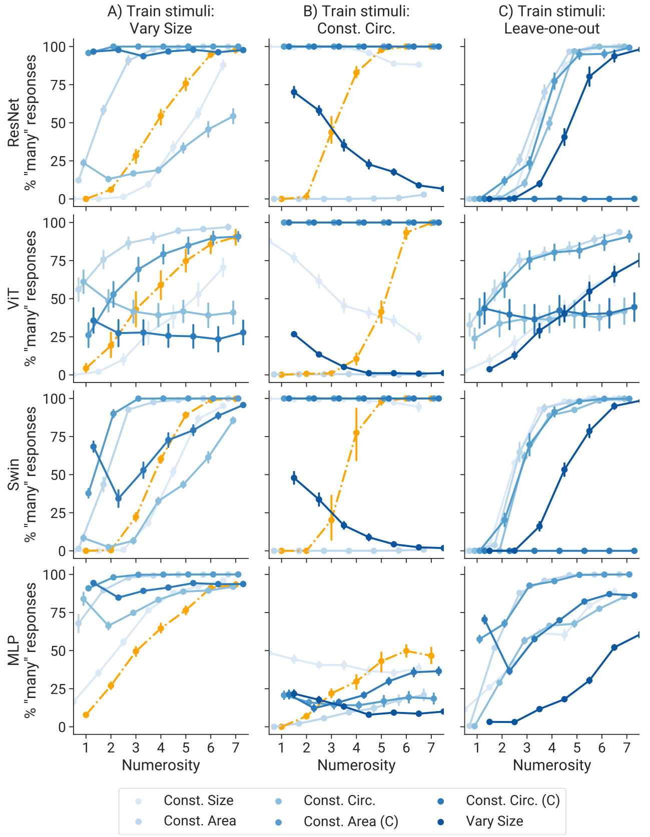

Figure 3: Psychometric curves in transfer experiments. Bars are standard errors of the mean. Models are represented across rows, and selected training datasets in columns. Shades of blue indicate the test stimulus, orange curves denote train and test on the same stimulus type.

Figures 3A) and 3B) show a selected subset of psychometric curves evaluated using such a cross-stimulus protocol, for models trained on *Vary Size* and *Constant Circumference*. Orange curves denote cases where a model is trained and tested on the same stimulus category (*i.e.,* the protocol from Exp. 1), and are included for reference, while blue curves are obtained when models are tested on datasets different from the ones they are trained on. We report results on *Vary Size* and *Constant Circumference* since they exhibit the most successful (*Vary Size*) and least successful (*Constant Circumference*) transfer cases, as defined by the expected qualititative characteristics of psychometric curves in Sec. The Numerical Bisection Task. Even for the best matching condition *Vary Size* (Fig. 3A), we observe a number of *transfer failures*, where a trained model shows poor transfer of number discrimination ability to a novel stimulus category, revealing a failure to abstract the number category.

An interesting case of transfer failures is evident in *all-or-none* responses, where models unanimously assign either *few* or *many* response to all numerosities. This has been observed across all models, and is particularly prominent with models trained on *Constant Circumference* (Fig. 3 B). In some cases, most notably with MLP and to a smaller extent with VIT, we also observe flatter curves with smaller slopes, resulting from frequent misclassification of *many* responses as *few* and vice versa. Finally, we also observe a new response pattern, an *inverted* psychometric curve, where small numerosities are overwhelmingly assigned label *many*, and the opposite for large numerosities. Fig. 3B) showcases that this pattern is consistent across models trained on *Constant Circumference*. We conjecture that this is due to models latching onto total white area during training, which is inversely correlated with numerosity in *Constant Circumference*.

Next, we consider an easier case of transfer, and examine if exposing models to more a diverse set of stimuli (as opposed to one type of stimulus) can help in learning a better representation of number categories; we train models on all but one stimulus type, and evaluate them on the hold-out stimulus type. Instead of 100 images per number category, a model is seeing 500 images per number category (*i.e.,* 100 images for a number for each of the 5 training stimulus types). As shown in Fig. 3C), we see that increasing data variability results in more curves that resemble the expected s-shaped curve, especially for RESNET and SWIN. However, even then, models failed to generalize to *Constant Circumference (contour)*, confirming again the difficulty of this stimulus category. Overall, we find that SWIN and RESNET produce representations that better match observed empirical data even in the more challenging transfer setting. We also observe that the training models on a variety of stimulus types help in generalising to new stimulus.

Finally, we examine if learned number representations form meaningful clusters; to answer this question, we do a forward pass on images from a given stimulus category for two models with lowest error rates (*i.e.,* RESNET and SWIN). For each image, we extract embeddings from the last dense layer of the model, prior to the 2-unit classification head. We use PCA followed by the t-SNE (Van der Maaten & Hinton, 2008) dimensionality reduction method to project high-dimensional embedding vectors (2,048 for RESNET and 768 for SWIN) into 2D space. In Fig. 4A) we show one selected example of such a projection, where individual points have been color-coded based on the numerosity of the stimulus image. First, we observe that embeddings cluster in groups based on number, with a greater cluster overlap for subsequent numbers. Second, we observe an ordering of clusters based on numerosities. This type of pattern is observed more often with embeddings from SWIN, compared to RESNET embeddings which generally result in less discernible clusters (with the exception of clusters for numerosities 1 and 2). Interestingly, based on visual inspection of the data, we do not find that more distinct projections suggest better performance on the task. For example, while clusters in Fig. 4A) seem to be discernible based on number, the model performs poorly when tested on *Constant Size*, possibly because the classifier does not discriminate based on the dimensions that are discriminable in the embeddings.

average with numerosity for both *Constant Size* and *Vary Size*.

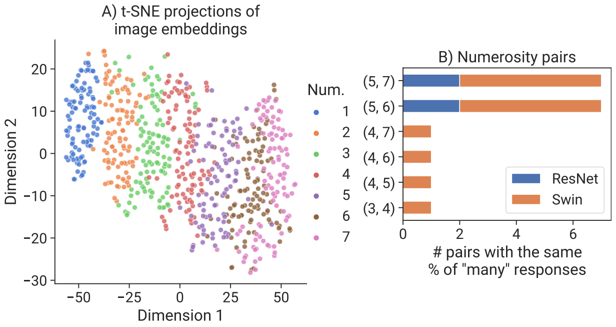

Figure 4: A) t-SNE projections of *Constant Size* image embeddings for Swin trained on *Constant Area*. B) Numerosities of image pairs for which there was no significant difference in the percentage of "many" responses.

Discriminability of Small vs. Large Numbers

Empirical data from humans and animals shows that in the numerical bisection task, it is consistently more difficult to distinguish larger numerosities from each other, compared to smaller ones (Almeida et al., 2007; Emmerton et al., 1997). This observation is likely to be related to a more general finding in numerical cognition that small numbers are processed differently than larger numbers (Dehaene, 1997; Revkin et al., 2008).

We examine whether similar observations can be made for our models' responses, using a measure of discriminability for different pairs of numbers. As an example, for a given model and a pair of numbers (such as 5 and 6), we statistically test if the mean percentage of *many* responses for images of one number (5) is the same as that of images of the other number (6). Intuitively, the two numbers are harder to discriminate if their mean percentage of *many* responses are the same. We consider models with smallest test error rates in Experiment 1 (*i.e.,* ResNet and Swin). For a given model and each stimulus type, we consider all possible number pairs (*i.e.,* all points on average psychometric curves) and perform the Tukey's HSD test for multiple pair-wise comparison of means (with family-wise error rate FWER=.05). This approach is based on the similar statistical tests used with pigeon responses in Emmerton et al. (1997).

In Fig. 4B), we show the breakdown of 18 cases (out of total 228 comparisons) where we failed to reject the null hypothesis across number pairs; intuitively, the models find it difficult to discriminate between these number pairs.[4] The figure shows that pairs at the higher end of the numerical range, such as (5, 7) and (5, 6) are frequently indistinguishable which is in contrast to the pairs on the lower end of the numerical scale. We conclude that similar to the empirical data, Swin and ResNet better distinguish numerosities at the lower end of the number range compared to those of the higher end of number ranges. Moreover, this effect is stronger among Swin responses compared to ResNet, suggesting that number representations learned by Swin are more discernible.

Discussion

Number discrimination is a core aspect of basic numerical competence in humans and animals. We investigate if recent, state-of-the-art neural networks used in computer vision exhibit the capacity of discriminating between small and large quantities. We evaluate these models on the numerical bisection task where models learn to categorize numerosity of sets of items, and we investigate their performance on novel stimuli and novel numerosities.

We find that ResNet and Swin, the two models with vision-specific inductive biases, achieve the smallest errors when categorizing novel stimuli as *few* or *many*. Psychometric curves of models trained on a wide range of stimuli, as well as those of models trained and tested on the same type of stimuli, often resemble the response curves of animals and humans on the same task. In addition, Swin responses are more discernible for smaller numbers compared to larger numbers, and its internal representations are structured in a way that reflects number category and order. Swin's predecessor, ViT, which is also a transformer-based model, albeit with a weak image-specific inductive bias, has errors on the task that are comparable to or even higher than the basic, substantially smaller MLP baseline. This is surprising considering that performance of ViT is within a few percentage points of Swin performance on different computer vision benchmarks (Liu et al., 2021).

Finally, when controlled for perceptual attributes (*e.g.,* keeping the total white area constant during training, but varying area during testing), most of these models show poor transfer of the number discrimination skill. This might mean that models latch onto features that, while correlated with number, are considered non-numerical in the literature (Honig & Stewart, 1989; Testolin et al., 2020). When analyzing the internal representations of number in Swin, we find that often, despite poor transfer, number representations are structured in an interpretable way. In other words, although these representations could in theory support number discrimination, we do not observe this in practice. One possible reason for poor transfer might be that the models are trained in a limited data regime, in contrast to humans and animals whose numerical cognition develops gradually in a rich environmental context, and who might be biologically predisposed to represent and process numerical quantities (Dehaene et al., 1998). Future work should explore whether pretraining models on larger and more diverse sets of images would result in a more transferable skill. Finally, we only investigate one specific task—the numerical bisection, and it remains to be explored whether our findings generalize across other perceptual domains.

Acknowledgments

The authors would like to thank Stephanie Chan for detailed feedback on this manuscript as well as other colleagues at

[4] From this plot we excluded 24 comparisons for number pairs (1, 2) and (6, 7) since means within these pairs are the same by the design of networks' training objective.

DeepMind for feedback and discussions that helped improve this work.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., … Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in neural information processing systems.*

Almeida, A., Arantes, J., & Machado, A. (2007, November). Numerosity discrimination in preschool children. *J. Exp. Anal. Behav.*, *88*(3), 339–354.

Anobile, G., Cicchini, G. M., & Burr, D. C. (2016). Number as a primary perceptual attribute: A review. *Perception*, *45*(1-2), 5–31.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Creatore, C., Sabathiel, S., & Solstad, T. (2021). Learning exact enumeration and approximate estimation in deep neural network models. *Cognition*, *215*, 104815. doi: https://doi.org/10.1016/j.cognition.2021.104815

Davis, H., & Memmott, J. (1982, November). Counting behavior in animals: A critical evaluation. *Psychol. Bull.*, *92*(3), 547–571.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics* [Book]. Oxford University Press New York.

Dehaene, S., & Changeux, J.-P. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of cognitive neuroscience*, *5*(4), 390–407.

Dehaene, S., Dehaene-Lambertz, G., & Cohen, L. (1998). Abstract representations of numbers in the animal and human brain. *Trends in neurosciences*, *21*(8), 355–361.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations.*

Droit-Volet, S., Clément, A., & Fayol, M. (2003, January). Time and number discrimination in a bisection task with a sequence of stimuli: a developmental approach. *J. Exp. Child Psychol.*, *84*(1), 63–76.

Emmerton, J., Lohmann, A., & Niemann, J. (1997, June). Pigeons' serial ordering of numerosity with visual arrays. *Anim. Learn. Behav.*, *25*(2), 234–244.

Fang, M., Zhou, Z., Chen, S., & McClelland, J. (2018). Can a recurrent neural network learn to count things? In *Cogsci.*

Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, *44*(1-2), 43–74.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673.

Gilmore, C. K., McCarthy, S. E., & Spelke, E. S. (2007). Symbolic arithmetic knowledge without instruction. *Nature*, *447*(7144), 589–591.

Halberda, J., Mazzocco, M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665–668.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Honig, W. K., & Stewart, K. E. (1989, June). Discrimination of relative numerosity by pigeons. *Anim. Learn. Behav.*, *17*(2), 134–146.

Jordan, K. E., & Brannon, E. M. (2006). Weber's law influences numerical representations in rhesus macaques (macaca mulatta). *Animal cognition*, *9*(3), 159–172.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., … Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 10012–10022).

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, *32*.

Meck, W. H., & Church, R. M. (1983, July). A mode control model of counting and timing processes. *J. Exp. Psychol. Anim. Behav. Process.*, *9*(3), 320–334.

Nieder, A. (2020). The adaptive value of numerical competence. *Trends in Ecology & Evolution*, *35*(7), 605–617.

Odic, D. (2018). Children's intuitive sense of number develops independently of their perception of area, density, length, and time. *Developmental Science*, *21*(2), e12533.

O'Shaughnessy, D. M., Gibson, E., & Piantadosi, S. T. (2021). The cultural origins of symbolic number. *Psychological review*.

Peterson, S. A., & Simon, T. J. (2000). Computational evidence for the subitizing phenomenon as an emergent property of the human cognitive architecture. *Cognitive Science*, *24*(1), 93–122.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008). Does subitizing reflect numerical estimation? *Psychological science*, *19*(6), 607–614.

Rodriguez, P., Wiles, J., & Elman, J. L. (1999). A recurrent neural network that learns to count. *Connection Science*, *11*(1), 5–40.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Testolin, A., Dolfi, S., Rochus, M., & Zorzi, M. (2020, June). Visual sense of number vs. sense of magnitude in humans and machines. *Sci. Rep.*, *10*(1), 10045.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Verguts, T., & Fias, W. (2004). Representation of number in animals and humans: A neural model. *Journal of cognitive neuroscience*, *16*(9), 1493–1504.

Zorzi, M., & Butterworth, B. (1999). A computational model of number comparison. *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society*, 772–777.