Computer Vision Modeling of the Development of Geometric and Numerical Concepts in Humans

Zekun Wang, Sashank Varma
Georgia Institute of Technology
School of Interactive Computing
Atlanta, GA 30332 USA
{zekun,varma}@gatech.edu

Abstract

Mathematical thinking is a fundamental aspect of human cognition. Cognitive scientists have investigated the mechanisms that underlie our ability to thinking geometrically and numerically, to take two prominent examples, and developmental scientists have documented the trajectories of these abilities over the lifespan. Prior research has shown that computer vision (CV) models trained on the unrelated task of image classification nevertheless learn latent representations of geometric and numerical concepts similar to those of adults. Building on this demonstrated cognitive alignment, the current study investigates whether CV models also show developmental alignment: whether their performance improvements across training to match the developmental progressions observed in children. In a detailed case study of the ResNet-50 model, we show that this is the case. For the case of geometry and topology, we find developmental alignment for some classes of concepts (Euclidean Geometry, Geometrical Figures, Metric Properties, Topology) but not others (Chiral Figures, Geometric Transformations, Symmetrical Figures). For the case of number, we find developmental alignment in the emergence of a human-like "mental number line" representation with experience. These findings show the promise of computer vision models for understanding the development of mathematical understanding in humans. They point the way to future research exploring additional model architectures and building larger benchmarks.

Introduction

Mathematical thinking is a fundamental aspect of human cognition, and as such has long been a target for AI researchers. Among the earliest AI programs were the Logic Theorist (Newell and Simon 1956), which proved theorems from *Principia Mathematica*, and Gelernter's geometry theorem prover (Gelernter 1959). There followed 60 years of steady progress on automating logico-mathematical reasoning, mostly within the symbolic paradigm. Over the past 10 years, rapid developments in ML have brought new successes to building systems that can reason mathematically. For example, in July 2025, the Gemini DeepThink model was able to meet the gold medal standard in the International Mathematical Olympiad (DeepMind 2025).

AI is both an engineering discipline and a scientific discipline. As the field develops more and more performant systems, we must also ask whether these systems represent mathematical concepts in the same ways people do. If so, then these systems can be brought into cognitive science as models of human mathematical thinking. In fact, this is increasingly the case. There is a long history in cognitive science of studies of the mental representations and processes by which people reason mathematically. Research over the past decade has shown that computer vision (CV) models and LLMs represent geometric and numerical concepts similarly to people (Shah et al. 2023; Stoianov and Zorzi 2012; Testolin, Zou, and McClelland 2020)

The vast majority of these studies have investigated the cognitive alignment between ML models and adult thinking. Here, we evaluate their potential developmental alignment: Does their improving mathematical performance across training match the developmental progressions observed in children? Researchers are only just beginning to move beyond the question of cognitive alignment to the question of developmental alignment (Frank and Goodman 2025; Shah, Bhardwaj, and Varma 2024; Warstadt and Bowman 2024). In this case study, we train a ResNet-50 model (He et al. 2015) on the ImageNet image dataset (Deng et al. 2009), measure as its sensitivity to geometric concepts grows and the precision of its number representations sharpens across checkpoints, and compare these progressions to those observed in children and adults across the lifespan.

Literature Review

The current study focuses on geometric and topological (GT) concepts and on number representations. This section reviews cognitive science studies of how adults understand geometry and number, and developmental science studies of the trajectories by which they come to this understanding. It also reviews investigations of whether ML models can capture these cognitive and developmental patterns. Although some of this work has been done with LLMs, we focus on CV models because this is the class of models explored in the current study.

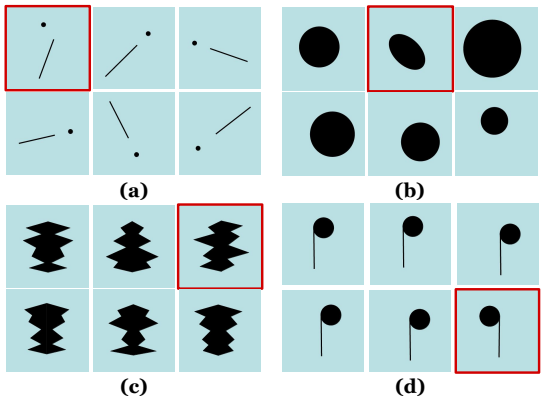

Figure 1: Sample stimuli for 4 GT concepts: (a) Euclidean Geometry - alignment of points in lines. (b) Geometrical Figures - circle. (c) Symmetrical Figures - vertical axis. (d) Chiral Figures - vertical axis. The odd-one-out is indicated by the red box.

Geometric and Topological Concepts The seminal study of how humans understand GT concepts is by Dehaene et al. (2006). They developed an odd-one-out task that tests people's sensitivity to 43 concepts, which themselves group into 7 classes: Topology, Euclidean Geometry, Geometric Figures, Symmetrical Figures, Chiral Figures, Metric Properties, and Geometrical Transformations. Figure 1 shows example stimuli for concepts from 4 classes. For each stimulus, the task is to judge which of the 6 images is the "odd one out". The images differ on multiple perceptual dimensions. Critically, 5 of the images embody the target concept whereas the 6th one does not. If people are sensitive to that concept, they will be above chance $(1/6)$ in selecting that image as the odd one out. Dehaene et al. (2006) administered this task to adults and children from the Mundurucu, an Amazon river valley group whose members have little or no formal schooling and are therefore unlikely to have received explicit instruction on these GT concepts. Nevertheless, they were above chance in selecting the odd-one-out for 37 of the 43 concepts (86%). The researchers also tested Western participants, finding that the children performed as well as the Mundurucu adults and children, and that the adults performed slightly better.[1] They interpreted the strong performance of the Mundurucu as evidence that people have *core knowledge* of GT concepts, which is to say they are part of the human endowment (Spelke and Kinzler 2007).[2]

Even from a strong core knowledge position, not all GT concepts need be part of a child's initial repertoire.[3] Rather, it is possible that some are available very early whereas others appear later, perhaps because they are learned through experience (Greenough, Black, and Wallace 1987). Izard and Spelke (2009) documented the developmental progression of GT concepts. In their Experiment 1, Western children ages 3-6 years old completed the odd-one-out task. The children showed above-chance sensitivity to 27 (63%) of the 43 concepts, suggesting that while some concepts might be part of core knowledge and available very early on, other concepts might be learned from experience in the world (including formal mathematics instruction). For example, the young children showed sensitivity to all 8 of the Euclidean Geometry concepts – but to none of the 8 Geometric Transformations concepts.

Recently, researchers have asked whether CV models are sensitive to GT concepts (Hsu, Wu, and Goodman 2022; Campbell et al. 2024). This is an interesting question because CV models are not trained to learn about mathematics. Rather, they are trained to accurately classify images. Thus, they can be understood as instantiating the view that (perceptual) development is mostly a matter of learning, which contrasts with the strict core knowledge view (Spelke and Kinzler 2007). This raises the question of whether, as a "side effect" of learning to classify images, CV models also become sensitive to GT concepts? If so, then the view of development as learning may be largely sufficient, and there may be less need to posit a role for core knowledge.[4]

Upadhyay et al. (2025) tested 5 CNN models on the odd-one-out task. The best performing model, ResNet-18 (He et al. 2015), showed sensitivity to 17 (40%) of the 43 GT concepts. This absolute level of performance was disappointing: though above chance (again, $1/6$), it was below that of the young children tested by Izard and Spelke (2009), who recall were sensitive to 27 (63%) of the 43 GT concepts. More promising was the correlation between the performance of the model and of the children at the level of the 7 classes of GT concepts, which was medium in size ($r = 0.52$, $p > .20$). This suggests that the model and the young children found the same classes of concepts relatively easy vs. difficult. Wang and Varma (2025) replicated these findings and extended them beyond CNNs to other model architectures: vision transformers and vision-language models. The vision transformer models they tested, ViT and DINOv2, achieved overall accuracies (47% and 49%, respectively) closer to the young children tested by Izard and Spelke (2009). Moreover, the correlations between the models and the young children across the 7 classes were exceptionally high: $r = 0.93$ and $r = 0.91$ ($ps < 0.01$), respectively.

Number Representation Cognitive science research has shown that people understand numbers by reference to a mental number line that is psychophysically scaled (Whalen, Gallistel, and Gelman 1999). Evidence for this representation comes from three effects, depicted in idealized form in the left panel of Figure 2. The *distance effect* is that when comparing which of two numbers n_1 and n_2 is greater, judgment time decreases linearly with the distance $|n_1 - n_2|$ between them (Moyer and Landauer 1967). This is consistent with

[1]The findings with Western adults have been replicated (e.g., Marupudi and Varma (2023)).

[2]It would make sense for evolution to deliver such an endowment given that the universe, and more locally the terrestrial environment, is governed by geometry and topology (Shepard 2001).

[3]We do not spring fully-formed from the brow of Zeus, like the goddess Athena.

[4]Of course, we can still ask whether the training of CV models and the (perceptual) development of children are analogous. We return to this question in the General Discussion.

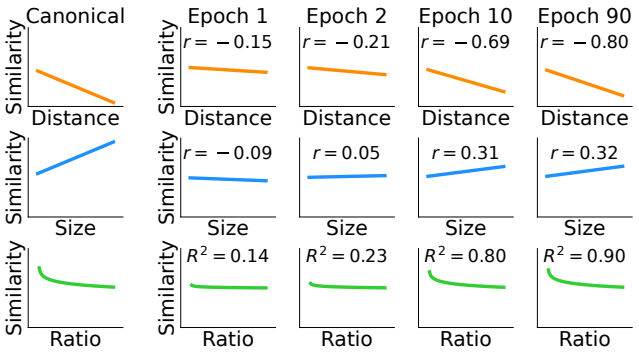

Figure 2: Idealized distance, size, and ratio effects observed in humans (left) and the emergence of these effects in ResNet-50 over training (right).

people locating the two numbers on their MNL and then discriminating which is one is "to the right" of the other. The farther apart they are, the easier the discrimination. The *size effect* is that when comparing two numbers, the greater their average size $(n_1+n_2)/2$, the slower the judgment time (Parkman 1971). For example, people take longer to compare 8 vs. 9 than 1 vs. 2 even though the distance is the same in both cases. This suggests that the distance between numbers is not constant, as in the conventional number line of mathematics, but psychophysically compressed, decreasing as numbers get larger. These two effects are combined in the ratio effect, which is that the time to compare two numbers decreases as the ratio of the larger to the smaller (i.e., $\max(n_1, n_2)/\min(n_1, n_2)$) increases (Gallistel and Gelman 1992). For example, people are very slow to compare 8 vs. 9 (ratio $= 1.125$), a bit faster to compare 1 vs. 2 (ratio $= 2.0$), and faster still to compare 1 vs. 9 (ratio $= 9.0$). People show these effects whether numbers are presented as digits (e.g., '3'), words (e.g., 'three'), or numerosities (e.g., '○ ○ ○') (Piazza et al. 2007). Importantly, the precision of the MNL improves over development, which can be seen in the sharpening of the distance, size, and ratio effects as children get older (Halberda, Mazzocco, and Feigenson 2008; Sekuler and Mierkiewicz 1977; Moore and Ashcraft 2015).

The presence of these effects very early in development, and even in other species, has led to the proposal that the MNL is "evolutionarily ancient" (Brannon and Terrace 1998; Nieder 2021). Alternatively, this representation might not be part of the human endowment, but rather learned "for free" through experience in the visual world. CV models can be used to test the sufficiency of this learning account. Stoianov and Zorzi (2012), in an early modeling study, trained a deep neural network on images depicting numerosities. The representations the model learned showed the distance and ratio effects, consistent with the model having learned a latent MNL. Subsequent studies with deep neural networks showed that this representation sharpens over training, paralleling its developmental trajectory in humans (Testolin, Zou, and McClelland 2020; Zorzi and Testolin 2017). More recent research utilizing conventional CV models – CNNs trained on ImageNet – has also found evidence of a latent MNL representation (Nasr, Viswanathan, and Nieder 2019).

In the closest prior study, Upadhyay and Varma (2023) evaluated the latent number representations of multiple pre-trained CNNs such as VGG19 (Simonyan and Zisserman 2015). They presented images of the numerosities $1-9$ and read off the vector representation on the final fully-connected layer of these models. VGG19 showed strong distance, size, and ratio effects, signaling that an MNL representation had been learned. They used multidimensional scaling (MDS) to reconstruct this representation, finding that it differed from the canonical MNL only in switching the positions of 1 and 2.

Research Questions

Previous research has established the sensitivity of CV models to GT concepts and has shown that CNN models possess latent number representations similar to the MNL of humans. With one notable exception (e.g., Testolin, Zou, and McClelland (2020)), this research has focused on the question of cognitive alignment, i.e., the correspondence of models to adult thinking. Here, we ask the question of developmental alignment:

1. Over training, does the sensitivity of ResNet-50 to GT concepts increase, and does this increase follow the developmental trajectory observed in people?
2. Over training, do the number representations of ResNet-50 increasingly show the distance, size, and ratio effect that signal an MNL representation, and does the precision of this representation improve according to the developmental trajectory observed in children?

Experiment 1

Experiment 1 investigated research question (1).

Method

Model and Training For this case study, we chose the ResNet-50 model (He et al. 2015) because Upadhyay et al. (2025) found that among the 5 CNNs they tested, it showed the greatest sensitivity to GT concepts and also moderate alignment with young children.[5] Furthermore, the architecture of CNNs maps closely to that of the human visual system, making them better candidates as cognitive (neuro)science models than other CV model architectures (Kriegeskorte 2015; Yamins and DiCarlo 2016).

In greater detail, our network followed the standard ResNet-50 configuration: a 7×7 conv (64 channels, stride 2) + BN/ReLU, 3×3 max-pool (stride 2), four residual stages with bottleneck blocks in the pattern $[3, 4, 6, 3]$ and output widths $[256, 512, 1024, 2048]$, global average pooling, and a 1,000-way fully connected classifier (about 25.6M parameters). We trained on ImageNet-1k (ILSVRC-2012) (Deng et al.

[5]The subsequent study by Wang and Varma (2025) found that the vision transformer models ViT (Dosovitskiy et al. 2021) and DINOv2 (Oquab et al. 2024) showed better overall performance and stronger developmental alignment than CNNs. However, we were unable to locate training checkpoints for either of these models and lacked the compute budget to train them ourselves.

2009) using the official train/validation split (1.28M/50k images). Training images were processed following the original implementation with `RandomResizedCrop` to 224×224 (scale $[0.08, 1.0]$, aspect ratio $[\frac{3}{4}, \frac{4}{3}]$), random horizontal flip (p=0.5), and per-channel normalization to ImageNet mean/std. Validation resized images to 256×256 and then 224×224 center-cropped with identical normalization. We optimized cross-entropy loss between the predicted class label and the actual labels with SGD, training for 90 epochs with global batch size 256, at an initial learning rate of 0.1 A step-scheduler was used to decrease learning rate by a factor of 0.1 every 30 epochs, ending training at a learning rate of 1×10^3. Runs use PyTorch on a single A40 (48 GB) GPU. We saved a full checkpoint (weights, optimizer/scheduler state, RNG) at the end of every epoch. Developmental analyses below use the sequence of checkpoints at saved epochs. Validation accuracy after training matches the standard ResNet-50 reference (top-1 ∼76%, top-5 ∼93%), confirming that our model is comparable to widely reported baselines and suitable for subsequent developmental alignment evaluations.

Design and Materials The stimuli were from Dehaene et al. (2006). As described above, there is one stimulus for each of 43 GT concepts (e.g., 'holes'); see Figure 1 for examples.[6] Each stimulus is composed of 6 images where 5 embody the GT concept and 1 does not. The task is to choose the 'odd one out'. The correct choice is the image that does *not* embody the GT concept, and so chance is $1/6$. The 43 GT concepts can be aggregated into 7 broader classes: Topology, Euclidean Geometry, Geometrical Figures, Symmetrical Figures, Chiral Figures, Metric Properties, and Geometrical Transformations. See Table 1 of the Supplementary Materials for a listing of all GT concepts and the classes to which they belong.

Human Data The human data were from Experiment 2 of Izard and Spelke (2009), which investigated the development of sensitivity to GT concepts across the lifespan. That study tested 400 Western participants ages $6-51$ years old. Most of the participants were children, adolescents, or young adults (i.e., 28 years old or younger); see the Supplementary Materials Figure 1 for a histogram of participant ages. Participants completed 2 practice trials followed by 43 experimental trials. On each trial, a stimulus was shown and participants clicked their choice of the odd-one-out image.

Procedure After each training epoch, we ran the model on the odd-one-out task, following the same method of Upadhyay et al. (2025) and Wang and Varma (2025). For each stimulus, each of the 6 images was first rescaled and cropped to 224×224 pixels. Each image was passed through the model and the representation before the final prediction layer collected as a vector of 2048 activations. Next, the cosine similarity between each pair of image vectors was computed. The model's choice of the odd-one-out image was the one with the lowest average cosine similarity to the other 5 images. We aggregated the model's performance to compute its overall accuracy and its accuracy for each of the 7 classes. These measures exactly parallel those computed for the human data.

[6]We thank Dr. Stanislas Dehaene for providing the stimulus images from this study. For further information about this dataset, please contact him directly.

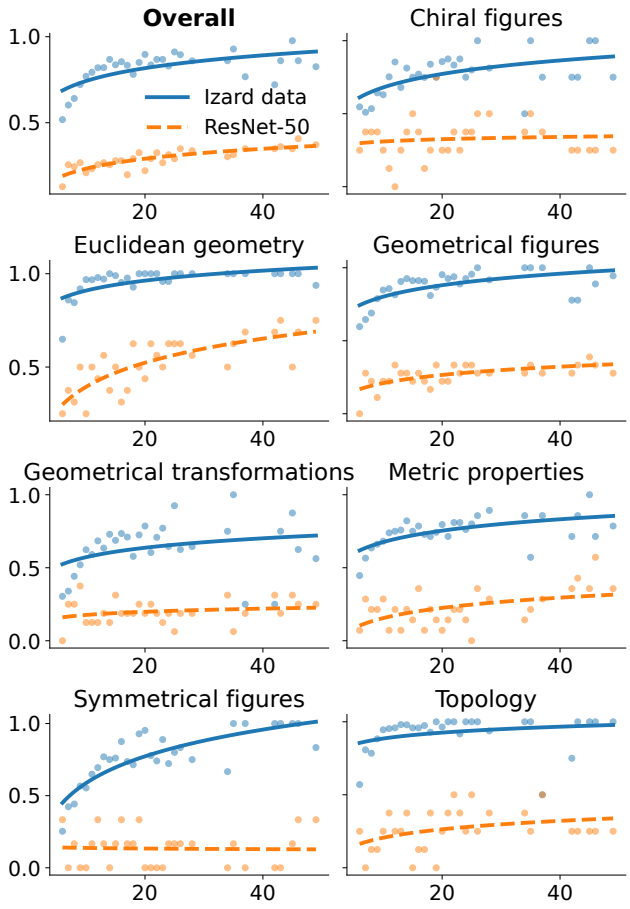

Figure 3: Overall accuracy for humans over development and for ResNet-50 over training (top-left panel), and for these data broken down for each of the 7 classes of GT concepts (remaining panels). x-axis: age (epochs); y-axis: accuracy.

Results

In the visualizations and analyses that follow, we mapped 2 epochs of model training to 1 year of human development. This was the natural mapping as the model was trained for 90 epochs and the age range of the sample was 45 years.

The top-left panel of Figure 3 shows the overall accuracy curves for humans over development and for ResNet-50 over training. Both humans and the model show improving performance with experience: the Pearson's correlation between them over years $6-51$ corresponding to epochs $2, 4, 6, \cdots, 90$ is $r = 0.50$ ($p < 0.01$). Because many human learning curves follow a power function (Laird, Rosenbloom, and Newell 1986; Logan 1988), we fit a power function to each set of data. This function offered a good account of the human data ($R^2 = 0.40$) and also of the model data ($R^2 = 0.66$), giving further evidence of developmental alignment. That said, humans decisively outperform the model.

The remaining panels of Figure 3 show the average ac-

curacy curves over development / training for each of the 7 classes. Humans show improving performance across development for all 7 classes, with the rate of improvement following a power function. By contrast, the model shows improving performance for only 4 of the 7 classes: Euclidean Geometry, Geometric Figures, Metric Properties, and Topology. (And for these 4 classes, model performance again lags human performance.) For the remaining 3 classes – Chiral Figures, Geometric Transformations, and Symmetrical Figures – the model's accuracy is both low and hardly improves across training. The associated correlations between human and model performance for each of the 7 classes as well as the fits of power functions can be found in Table 2 of the Supplementary Materials.

Discussion

Experiment 1 investigated research question (1): Whether ResNet-50's growing sensitivity to GT concepts over training matches the trajectories observed in humans over development. The model's overall performance on the odd-one-out task improves with training according to a power function, matching the trajectory observed in humans – although the model's absolute level of performance is lower. At a finer-grain level, the model shows growing sensitivity for 4 of the 7 classes. This suggests that Euclidean Geometry, Geometric Figures, Metric Properties, and Topology concepts might come "for free" when learning to perceive the visual world, and need not be entirely located within core knowledge. By contrast, for Chiral Figures, Geometric Transformations, and Symmetrical Figures concepts, the model shows almost no improvement over training. This stands in contrast to humans, who show improved sensitivity over development. This is evidence that these concepts do not come "for free", and instead might be part of core knowledge or be learned through explicit mathematics instruction.

Experiment 2

Experiment 2 investigated research question (2).

Method

Model and Training Same as Experiment 1.

Design and Materials The stimuli were from Upadhyay and Varma (2023). Each is a 720×720-pixel image showing a numerosity of $1-9$ items. The stimuli are organized into 6 sets that vary in which perceptual variables are controlled, which are varied parametrically, and which are allowed to vary randomly. The stimulus sets are intended to be progressively more difficult for models, to enable titration of their sensitivity to numerosity (over perceptual variables).

1. The items are black circles randomly placed on a white background. For a given area A, the total area of each numerosity (i.e., the number of black pixels) is controlled to be A. Thus, a stimulus with numerosity 1 and another with numerosity 9 each have A black pixels. This prohibits using this perceptual feature (total area) as a proxy for numerosity. The total area is parametrically varied across five levels A_1 – A_5 corresponding to 103 – 518 black pixels, defining five subsets of images.

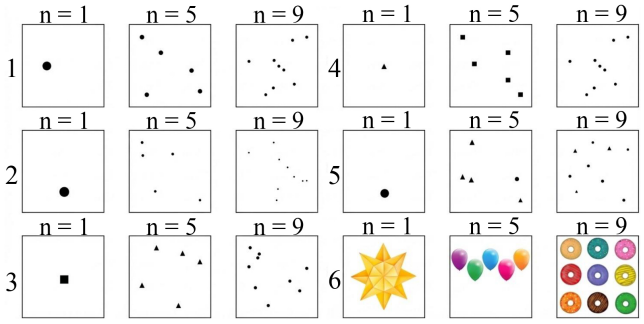

Figure 4: Example stimuli for the stimulus sets $1-6$.

2. Like (1) but the total circumference C is controlled, so that this feature cannot be used as a proxy for numerosity. The total circumference is parametrically varied across five levels C_1 – C_5 corresponding to 100 – 300 pixels, defining five subsets of images.
3. Like (2) but the items of the two numerosities are randomly varied, e.g., circles in one image and squares in another. This enables testing generalization across shapes.
4. Like (3) but the total areas of the two numerosities is randomly varied, i.e., one having area A_i and the other area A_j ($i \neq j$). This enables generalization across both shapes and areas.
5. 'Anything goes': Like (4) except the individual items of each numerosity are randomly varied so that each is a mixture of circles, squares and triangles of different areas. This enables further generalization of the findings.
6. Naturally occurring numerosities found using Google Search (and manually verified). These are mostly stylized like clip art. These items vary on many perceptual dimensions (e.g., shape, size, drawing style, color, etc.), enabling further generalization of the findings.

See Figure 4 for example stimuli from each set, and the Supplementary Materials for further details on their construction.

Procedure After each training epoch, we evaluated the model's distance, size, and ratio effects for each stimulus set. Recall these effects are collectively evidence for an MNL representation. Given that the numerosities are in the range $1-9$, there are $(9 \times 8)/2 = 36$ pairs of numerosities n_1 and n_2 such that $n_1 \neq n_2$ (i.e., so that one is more numerous than the other).

For each stimulus set (and for each level of total area A or total circumference C, if relevant), for each of the 36 comparisons, we randomly sampled stimuli of numerosity n_1 and n_2. We passed each through the model and captured the vector representation before the final prediction layer. We then computed the cosine similarity between the two vectors. We made the following linking hypothesis to map model performance to human performance: the less similar the vectors, the more discriminable the corresponding numerosities, and thus the faster the predicted time to judge which one is the greater numerosity. This is the same linking hypothesis that has been used in prior studies of numerical alignment between humans and CV models (Upadhyay and

Varma 2023) and LLMs (Shah et al. 2023). The three effects were computed as follows:

- distance effect: The correlation between the similarity of the vectors and the distance between the numerosities $|n_1 - n_2|$. A negative correlation indicates a human-like distance effect.
- size effect: The correlation between the similarity of the vectors and the average size of the numerosities $(n_1 + n_2)/2$. A positive correlation indicates a human-like size effect.
- ratio effect: The R^2 of fitting a negative exponential function predicting the similarity of the vectors by the ratio of the larger numerosity to the smaller: $\max(n_1, n_2)/\min(n_1, n_2)$. A value closer to 1 indicates a human-like ratio effect. Canonical distance, size, and ratio effects are shown in the left panel of Figure 2.

Results

Research question (2) asks if the number representations of ResNet-50 develop over training along the same trajectory as the MNL of humans sharpens over development. To visualize what this would mean, the right panel of Figure 2 plots the distance, size, and ratio effects after epochs 1, 2, 10, and 90 of training. (We chose these epochs because the model rapidly learns at the earlier training stages.) We see that the effects are absent early in training, signaling the absence of an MNL representation. However, over training, these effects manifest. Thus, as a "side effect" of learning to classify images, the model learns a human-like representation of number.

At a more detailed level, we can plot the trajectory of these effects over all 90 epochs. This is shown in Figure 5 – the correlations for the distance and size effects and the R^2 for the ratio effect. We see that the distance effect is robust: it appears early in training, follows the canonical functional form (i.e., a negative correlation), and holds for all but the most varied stimulus sets (1 and 6). The ratio effect is also robust, following the canonical functional form (i.e., the R^2 is high) for all but the most varied stimulus sets (5 and 6). By contrast, the size effect is smaller in size, with correlations positive (as predicted) but closer to 0 than 1. Curiously, the size effect is weakest in the 'easiest' stimulus sets: 1 (equal-area circles) and 2 (equal-circumference circles).

We conducted a growth curve analysis of the developmental trajectories in Figure 5. Specifically, for each stimulus set, we fit a power function to each of the three effects. We refer the reader to Table 3 of the Supplementary Materials for the fits of power functions. The overall pattern is for developmentally plausible growth of the distance and ratio effects, with a power function characterizing the improvement of the (negative) correlation and the R^2 fit value, respectively, over training epochs. This holds for all but the most varied stimulus set (6). By contrast, the growth of the size effect is less human-like, with the power function offering a generally worse account of the improvement of the (positive) correlation over training epochs.

Finally, we followed Upadhyay and Varma (2023) and reconstructed the latent number line representation of the model at each epoch. Specifically, for stiumulus set (1), we formed a

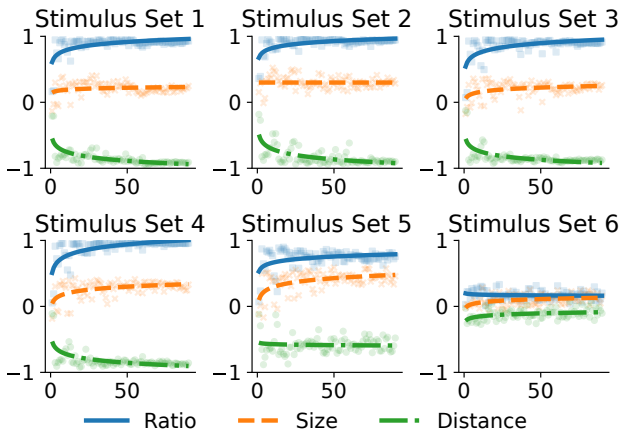

Figure 5: The distance effect correlation (green; expected to be close to -1), size effect correlation (orange; expected to be close to 1), and ratio effect R^2 (blue; fit of the negative exponential function, expected to be close to 1) of ResNet-50 over training. x-axis: epochs; y-axis: correlation or R^2.

9×9 matrix where each entry is the cosine similarity between the vector representations of the corresponding numerosities before the final prediction layer. We submitted these pairwise similarities to MDS and requested a 1D solution, which we interpret as the model's latent number line representation at that point in training. Figure 6 plots these for epochs 1, 2, 10, and 90. (Again, we chose these epochs because the model learns rapidly.) Over training, this representation comes to resemble the canonical MNL of humans, further showing the model's developmental alignment.

Discussion

Experiment 2 investigates research question (2): Whether, over training, the number representations of ResNet-50 increasingly show the distance, size, and ratio effects that signal an MNL representation? This was the case. Across a range of stimulus formats, the model showed the distance and ratio effects early in training, and these effects only strengthened over time. There was less evidence for the orderly emergence of the size effect, and its generalization was lower to more varied stimulus presentation formats. Finally, a reconstruction of the model's number line representation over training shows the increasing sharpening of its MNL, further evidencing its developmental alignment. These findings support the proposal that visual experience in the world may deliver an MNL representation "for free", and there may be less need to posit that it's part of core knowledge.

General Discussion

Prior studies have used computer vision models to investigate mathematical thinking (Boccato, Testolin, and Zorzi 2021; Kim et al. 2021; Nasr, Viswanathan, and Nieder 2019; Stoianov and Zorzi 2012; Testolin, Zou, and McClelland 2020; Upadhyay and Varma 2023; Upadhyay et al. 2025; Wang and Varma 2025; Zorzi and Testolin 2017). Most have

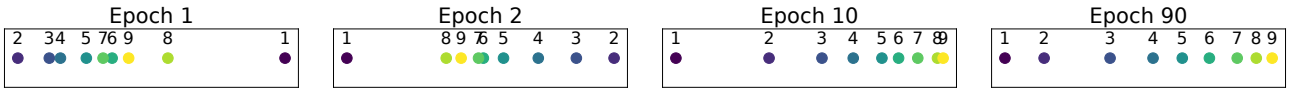

Figure 6: Reconstructed number line representations of ResNet-50 over training showing the sharpening of its MNL.

focused on adult cognition, with only Testolin, Zou, and McClelland (2020) exploring the question of cognitive development. However, this study suffers from several limitations: It utilized a custom deep neural network as opposed to a standard CNN or vision transformer architecture, it used a custom 'layer-wise' training procedure rather than a standard procedure, and it used a custom training dataset of abstracted stimuli rather than naturalistic images. The current study takes an important step beyond this earlier work.

We asked whether a standard CV model architecture trained on a standard image dataset shows human-like trajectories in the growth of GT concept sensitivity and number representation precision. We chose the ResNet-50 model for this case study because prior research has demonstrated its cognitive alignment with how adults represent geometric concepts (Upadhyay et al. 2025). We trained it on the ImageNet image dataset (Deng et al. 2009) and saved model checkpoints along the way. Experiment 1 found increasing sensitivity to four classes of GT concepts over training – Euclidean Geometry, Geometric Figures, Metric Properties, and Topology – mimicking the trajectory observed in humans (albeit with lower overall performance). However, there was no improvement with training for the three other classes – Chiral Figures, Geometric Transformations, and Symmetrical Figures. Experiment 2 probed the development of number representations. That humans understand numbers by reference to an MNL is evidenced by the distance, size, and ratio effects. Moreover, these effects sharpen over development, signaling an increase in the precision of this representation. This was also the case for ResNet-50 over the course of training, most strongly for the distance and ratio effects and for stimulus sets 1 – 4.

An important question for developmental science is: *what develops?* Experiment 1 gives suggestive but no definitive answer for GT concepts: some GT concepts might come "for free" from learning to perceive the world, whereas other concepts appear not to be so easily learnable. This might signal that these latter concepts are part of the child's core knowledge' (Spelke and Kinzler 2007), and thus the mind/brain does not have to be architected to learn them from experience. (Another interpretation is that they do not belong to core knowledge either, and instead must be learned from supervised mathematics instructions.) Experiment 2 gives a clearer answer to the question: what develops is the model's latent number line representation, which becomes increasingly canonical over training; see Figure 6. This is the same "mechanism of change" proposed by mathematical development researchers (Halberda, Mazzocco, and Feigenson 2008; Sekuler and Mierkiewicz 1977; Moore and Ashcraft 2015).

Together, these results show the continuing promise of computer vision models for advancing developmental science. However, for this potential to be realized, several limitations must overcome.

The first limitation concerns the assumption that training on the image classification task is a valid proxy for humans learning to perceive the visual world. This is almost certainly not the case. Vision is useful for object recognition, to be sure, but also for many other functions, such as tracking the movement of objects in space (visual attention) (Corbetta and Shulman 2002; Szczepanski et al. 2013) and reasoning about visuospatial problems (e.g., mental rotation) (Zacks 2008; Tomasino and Gremese 2016). This gap presents an opportunity. The current study failed to find evidence of growing sensitivity to three classes of GT concepts over training. Perhaps this failure reflects the limits of the image classification task. Future work should explore training CV models on a range of tasks more representative of the range of tasks the human visual system can perform. It may be that additional classes of GT concepts are learned "for free" under such an expanded training procedure.

A second limitation is the limited nature of the mathematical measures used. In testing sensitivity to GT concepts, each concept was represented by only one stimulus. It is possible that the 5 images that embody a concept also shared other perceptual properties which are not present in the odd-one-out image, and that these properties instead drove model performance. A stronger benchmark would include many more stimuli for each concept. Experiment 2 used a broader range of stimuli (6 sets) to evaluate the development of number representations over training. The distance and ratio effects were weakest for the most varied stimulus sets (5 and 6) that used the most "naturalistic" stimuli, including clip art images from a Google Image search. Future work should use even more visually complex images, such as stimuli from the MS COCO (Lin et al. 2015) and CLEVR (Johnson et al. 2016) datasets, to further test the robustness and generalization of the latent number representation learned by CV models.

A third limitation is that this case study explored only one model architecture, a CNN trained on one image dataset. It likely underestimates the potential developmental alignment of CV models. Future research should explore a variety of model architectures trained on a range of datasets. For example, Wang and Varma (2025) found that vision transformers like ViT and DINOv2 achieve higher overall accuracy than CNNs on the odd-one-out task for GT concepts, closer to that of young children. At the finer grain of the 7 classes, the correlations between these models and the young children were exceptionally high ($r > 0.90$). It is possible that over training, vision transformer models may show strong alignment to the developmental trajectories of children.

References

Boccato, T.; Testolin, A.; and Zorzi, M. 2021. Learning Numerosity Representations with Transformers: Number Generation Tasks and Out-of-Distribution Generalization. *Entropy*, 23(7): 857.

Brannon, E. M.; and Terrace, H. S. 1998. Ordering of the numerosities 1 to 9 by monkeys. *Science*, 282(5389): 746–749.

Campbell, D.; Kumar, S.; Giallanza, T.; Griffiths, T. L.; and Cohen, J. D. 2024. Human-Like Geometric Abstraction in Large Pre-trained Neural Networks. arXiv:2402.04203.

Corbetta, M.; and Shulman, G. L. 2002. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3): 201–215.

DeepMind. 2025. Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad — deepmind.google. https://deepmind.google/discover/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad/. [Accessed 31-07-2025].

Dehaene, S.; Izard, V.; Pica, P.; and Spelke, E. 2006. Core Knowledge of Geometry in an Amazonian Indigene Group. *Science*, 311(5759): 381–384.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Frank, M. C.; and Goodman, N. D. 2025. Cognitive Modeling Using Artificial Intelligence. *Annual Review of Psychology*.

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

Gelernter, H. 1959. A note on syntactic symmetry and the manipulation of formal systems by machine. *Information and Control*, 2(1): 80–89.

Greenough, W. T.; Black, J. E.; and Wallace, C. S. 1987. Experience and Brain Development. *Child Development*, 58(3): 539.

Halberda, J.; Mazzocco, M. M. M.; and 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.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385.

Hsu, J.; Wu, J.; and Goodman, N. D. 2022. Geoclidean: Few-Shot Generalization in Euclidean Geometry. arXiv:2211.16663.

Izard, V.; and Spelke, E. S. 2009. Development of sensitivity to geometry in visual forms. *Hum. Evol.*, 23(3): 213–248.

Johnson, J.; Hariharan, B.; van der Maaten, L.; Fei-Fei, L.; Zitnick, C. L.; and Girshick, R. 2016. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. arXiv:1612.06890.

Kim, G.; Jang, J.; Baek, S.; Song, M.; and Paik, S.-B. 2021. Visual number sense in untrained deep neural networks. *Science Advances*, 7(1).

Kriegeskorte, N. 2015. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, 1(1): 417–446.

Laird, J. E.; Rosenbloom, P. S.; and Newell, A. 1986. Chunking in Soar: The anatomy of a general learning mechanism. *Machine Learning*, 1(1): 11–46.

Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312.

Logan, G. D. 1988. Toward an instance theory of automatization. *Psychological Review*, 95(4): 492–527.

Marupudi, V.; and Varma, S. 2023. Graded human sensitivity to geometric and topological concepts. *Cognition*, 232: 105331.

Moore, A. M.; and Ashcraft, M. H. 2015. Children's mathematical performance: five cognitive tasks across five grades. *J. Exp. Child Psychol.*, 135: 1–24.

Moyer, R. S.; and Landauer, T. K. 1967. Time required for Judgements of Numerical Inequality. *Nature*, 215(5109): 1519–1520.

Nasr, K.; Viswanathan, P.; and Nieder, A. 2019. Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Sci. Adv.*, 5(5): eaav7903.

Newell, A.; and Simon, H. 1956. The logic theory machine–A complex information processing system. *IRE Transactions on Information Theory*, 2(3): 61–79.

Nieder, A. 2021. The evolutionary history of brains for numbers. *Trends Cogn. Sci.*, 25(7): 608–621.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193.

Parkman, J. M. 1971. Temporal aspects of digit and letter inequality judgments. *Journal of Experimental Psychology*, 91(2): 191–205.

Piazza, M.; Pinel, P.; Le Bihan, D.; and Dehaene, S. 2007. A Magnitude Code Common to Numerosities and Number Symbols in Human Intraparietal Cortex. *Neuron*, 53(2): 293–305.

Sekuler, R.; and Mierkiewicz, D. 1977. Children's Judgments of Numerical Inequality. *Child Development*, 48(2): 630.

Shah, R.; Marupudi, V.; Koenen, R.; Bhardwaj, K.; and Varma, S. 2023. Numeric Magnitude Comparison Effects in Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2023*, 6147–6161. Association for Computational Linguistics.

Shah, R. S.; Bhardwaj, K.; and Varma, S. 2024. Development of Cognitive Intelligence in Pre-trained Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 9632–9657. Miami, Florida, USA: Association for Computational Linguistics.

Shepard, R. N. 2001. Perceptual-cognitive universals as reflections of the world. *Behavioral and Brain Sciences*, 24(4): 581–601.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556.

Spelke, E. S.; and Kinzler, K. D. 2007. Core knowledge. *Developmental Science*, 10(1): 89–96.

Stoianov, I.; and Zorzi, M. 2012. Emergence of a "visual number sense" in hierarchical generative models. *Nature Neuroscience*, 15(2): 194–196.

Szczepanski, S. M.; Pinsk, M. A.; Douglas, M. M.; Kastner, S.; and Saalmann, Y. B. 2013. Functional and structural architecture of the human dorsal frontoparietal attention network. *Proceedings of the National Academy of Sciences*, 110(39): 15806–15811.

Testolin, A.; Zou, W. Y.; and McClelland, J. L. 2020. Numerosity discrimination in deep neural networks: Initial competence, developmental refinement and experience statistics. *Developmental Science*, 23(5).

Tomasino, B.; and Gremese, M. 2016. Effects of Stimulus Type and Strategy on Mental Rotation Network: An Activation Likelihood Estimation Meta-Analysis. *Frontiers in Human Neuroscience*, 9.

Upadhyay, N.; Marupudi, V.; Varma, K.; and Varma, S. 2025. Alignment of CNN and Human Judgments of Geometric and Topological Concepts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2): 1556–1564.

Upadhyay, N.; and Varma, S. 2023. CNN models' sensitivity to numerosity concepts. In *The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS'23*.

Wang, Z.; and Varma, S. 2025. Computer Vision Models Show Human-Like Sensitivity to Geometric and Topological Concepts. arXiv:2505.13281.

Warstadt, A.; and Bowman, S. R. 2024. What Artificial Neural Networks Can Tell Us About Human Language Acquisition. arXiv:2208.07998.

Whalen, J.; Gallistel, C.; and Gelman, R. 1999. Nonverbal Counting in Humans: The Psychophysics of Number Representation. *Psychological Science*, 10(2): 130–137.

Yamins, D. L. K.; and DiCarlo, J. J. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.*, 19(3): 356–365.

Zacks, J. M. 2008. Neuroimaging Studies of Mental Rotation: A Meta-analysis and Review. *Journal of Cognitive Neuroscience*, 20(1): 1–19.

Zorzi, M.; and Testolin, A. 2017. An emergentist perspective on the origin of number sense. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 373(1740): 20170043.