# Examining Infant Relation Categorization Through Deep Neural Networks

**Guy Davidson (guy.davidson@nyu.edu)**
Center for Data Science
New York University

**Brenden M. Lake (brenden@nyu.edu)**
Department of Psychology and Center for Data Science
New York University

## Abstract

Categorizing spatial relations is central to the development of visual understanding and spatial cognition, with roots in the first few months of life. Quinn (2003) reviews two findings in infant relation categorization: categorizing one object as above/below another precedes categorizing an object as between other objects, and categorizing relations over specific objects predates abstract relations over varying objects. We model these phenomena with deep neural networks, including contemporary architectures specialized for relational learning and vision models pretrained on baby headcam footage (Sullivan et al., 2020). Across two computational experiments, we can account for most of the developmental findings, suggesting these neural network models are useful for studying the computational mechanisms of infant categorization.

**Keywords:** neural networks, spatial categorization, infant relation learning, developmental computational modeling

Relations are critical to human reasoning capacities (Goodwin and Johnson-Laird, 2005), and our understanding of the visual world around is mediated by spatial relations, as they help distinguish individual objects and combine them in order to understand visual scenes (Piaget, 1954; Johnson, 2010). Additionally, both relational learning (Newcombe and Huttenlocher, 2007) and analogical reasoning (Yuan et al., 2017) appear crucial to the development of spatial cognition, which guides infants' budding understanding of the world around them. Despite the importance of relations, little computational work has examined how infants could learn to categorize spatial relations, and why some categories are acquired before others over the course of development. This is the goal of our current article.

Relational learning and reasoning have received substantial recent attention in the artificial intelligence literature (see Battaglia et al., 2018, for a review). Santoro et al. (2017) and Shanahan et al. (2019) offer novel neural network architectures designed for relational reasoning, while Barrett et al. (2018) and Teodorescu et al. (2020) offer diagnostic task paradigms. Other literature focuses on applications, such as Hamrick et al. (2018) and Kipf et al. (2018), exploring physical property inference through relational reasoning. We turn this recent attention toward modeling infant categorization of spatial relations to better understand the computational basis of these early-emerging abilities. We evaluate a wide variety of neural network architectures trained on both synthetic and real-world datasets, comparing their performance with key findings from the developmental literature. Previous work used very simple connectionist networks to model aspects of infant categorization (Mareschal et al., 2000; French et al., 2004) and spatial language (Regier, 1995). Our contribution is to evaluate the latest generation of architectures, especially those specialized for processing spatial relations. Next, we will describe the developmental findings we model in our computational experiments.

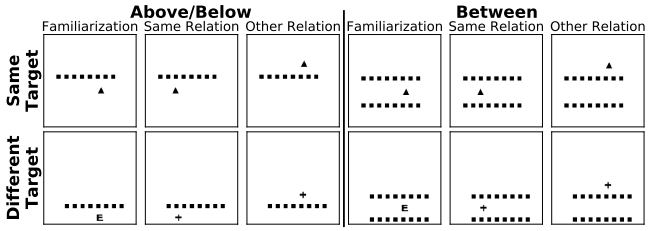


**Figure 1: Example Stimuli.** In each triplet, the left stimulus is the familiarization example, the middle is the same relation test, and the right stimulus is the other relation test. Top: identical target objects. Bottom: different target objects with alternative reference bar.

Quinn (2003) surveys the development of infant relation categorization, including two primary findings. The first finding is that, by 3-4 months of age, infants can categorize "above versus below" (or "left versus right", Quinn, 2004), although they fail to categorize "between" (Figure 1; top row). By 6-7 months, infants can also categorize "between." In a representative experiment, Quinn (1994) familiarized infants with several stimuli, all containing a dot either above or below a horizontal bar (Figure 1; Familiarization). After familiarization, infants were presented with a novel category preference test, finding that infants look longer at a stimulus with the dot on the other side of the bar (Figure 1; Other relation) compared to a new location on the same side (Figure 1; Same relation).

The second finding is that infants categorize spatial relations comprised of specific objects before categorizing the same relations composed of varying objects. Quinn et al. (1996; 2003) replicate the previous experiments except that the target object varies between familiarization and test (Figure 1; bottom row). In both cases, changing the target object requires the infants to be older to show the same novelty preference—from 3-4 months to 6-7 months for above versus below, and from 6-7 months to 9-10 months for between versus outside.

When infants discriminate between categories in a laboratory study, it is often unclear whether these abilities reflect top-down processing of categories acquired outside the lab, or bottom-up processing of categories developed during the familiarization phase (Thelen and Smith, 1994; Murphy, 2002, ch. 9; French et al., 2004, Newcombe et al., 2005). In this paper, we examine neural networks as a tool for studying both types of processing. In experiment 1, we report a supervised learning paradigm examining learning different relations from simplified object vectors. We view this as analogous to the first possibility, evaluating the difficulty of learning relation concepts from numerous varied examples, as infants might acquire these categories over an extended period outside the

lab. We compare five neural network architectures, instantiating different inductive biases, to observe which show similar patterns to the results discussed in the first finding above. In experiment 2, we appraise the second possibility, evaluating whether objects arranged in the same spatial relation are encoded more similarly than objects that are not, based entirely on general purpose, high-level perceptual features. To do so, we utilize learned features from large-scale computer vision architectures as a proxy for prior visual experience, pretrained on either a developmentally-realistic visual corpus or a popular computer vision benchmark, neither of which explicitly requires relational categorization.[1] Our results show that both experiments account for the primary findings, suggesting that neural networks can serve as useful models for both types of categorization processes. The variation between training architectures (both experiments) and pretraining methods (experiment 2) suggests that some networks better model the development of relation learning, and gaps in the results highlight promising approaches for developing more comprehensive models of categorization in infancy.

## Experiment 1: Relations from Scratch

In this experiment, we model the first finding discussed, that infants acquire the capacity to represent "above or below" (a target object relative to a single reference object) before they develop the ability to represent "between" (a target relative to two references). In two studies (Quinn, 1994; Quinn et al., 1996), 3-4 months old infants familiarized with stimuli depicting a single relation (either above or below) exhibit a looking-time preference to a stimulus showing the opposite relation, compared to a new stimulus showing the familiarized relation. Quinn et al. (1999) followed up on those experiments, using examples of a target object between two reference objects, using both horizontal and vertical reference objects. 3-4 months old infants did not display a preference towards test stimuli containing an object outside the references, but infants 6-7 months old did. The ability to reason relative to two reference objects develops after the ability to reason relative to a single reference, consistent with the notion that infants first encode with respect to a single landmark, and later encode in a "local spatial framework" (Huttenlocher and Newcombe, 1984).

### Methodology

Our simulations evaluate the relative ease of learning two different classes of relations, both cast as binary classification problems: *above/below* (learning to classify above versus below), and *between* (learning to classify between versus outside).

**Objects.** To model relation learning independently from learning to represent objects, we provide the models with minimal object representations as inputs. Each object is represented as a vector of length 4, with integer $x$ and $y$ positions, and a one-hot encoding marking the object as target or reference (Figure 2 bottom). The objects are implicitly understood to be occupying a 1x1 unit square. The reference objects, which we take to be 9 units long, are represented as a collection of 9 adjacent identically-sized objects. We also explored an alternative representation that treats the reference bar as a single object, where each object vector has an additional integer dimension specifying its length (as all objects we use have a height of 1 unit, we omit a height dimension). Results with the alternative representation were qualitatively similar, even though the task is easier (as the models receive fewer object vectors as their input), so we focus on describing the results with the first representation (without the length dimension).

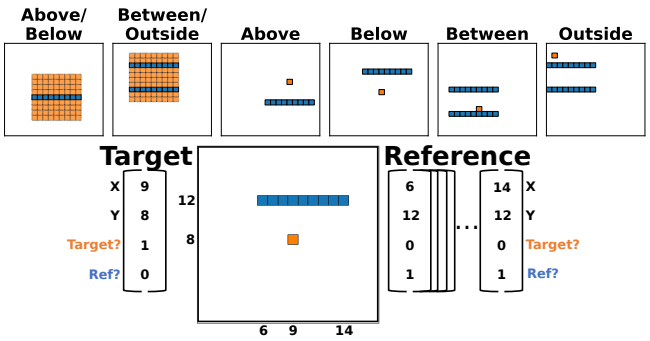


**Figure 2: Experiment 1 Stimuli. Top:** Left two panels: a sample location of the reference object(s) (in blue), with the entire grid of possible target object locations visualized (in orange). Middle two panels: example *above/below* stimuli. Right two panels: example *between* stimuli. **Bottom:** the vector object representations associated with the Below example—as the models receive only these vectors, the choice colors and shapes here is arbitrary. We do not mark which coordinate is X and which is Y, so the models are agnostic to this fact (and above/below is identical to left/right), other than the CNN model, which receives a spatial input. The borders signify each object vector, so the blue reference object is comprised of nine vectors.

**Dataset Generation.** Figure 2 visualizes stimuli from the different relation categories. To create stimuli, we sample locations for the reference object (series of blue cells) and then sample the target object's location uniformly from the 'target grid' above and below the reference object(s). In the *above/below* condition, we split the eight rows of the target grid evenly between above and below. In the *between* condition, we split the locations evenly between the between and outside relations. We only consider cases where the target object occupies the same horizontal space as the reference object(s), avoiding having the target object off to the side. To create training and test sets, we randomly split the reference object locations (in the large canvas, 90% training, 10% test) and the target object locations (relative to the reference object, 80% training, 20% test). We then set aside 10% of the training set as a validation set. This process creates a maximal training set of 3628 examples, a validation set of 404 examples, and a test set of 1800 examples. We also evaluate models trained on randomly sampled subsets of the training sets, using 8, 32, 128, 512, 1024, or 2048 items.

[1] If the pretrained models demonstrate categorical perception, it's possible that they acquired perceptual features akin to abstract relational categories. A more likely possibility is that they acquired perceptual features that implicitly promote relational similarity more than the alternative, of acquiring perceptual features that promote relational opposition.

**Architectures.** We evaluated five different neural networks, each incorporating a distinct inductive bias. To the extent possible, the architectures were chosen to gracefully handle varying numbers of objects present in a scene. Other than the convolutional neural network, all models begin with an object-wise embedding function, a single layer with ReLU activations. We denote the input collection of objects as $O = \{o_1, ..., o_N\}$, the embedding function as $e_\omega$, and the embedded objects as $E = e_\omega(O) : \{e_i = e_\omega(o_i)\}$. All models have two softmax output units (the two classes learned), and are trained using the cross-entropy loss to maximize the probability of the correct class.

*'Bag of objects' MLP:* this architecture is the simplest we could conceive of that would be invariant to the number of objects. It treats the embedded vectors as a single vector by taking their mean, and passes it into a standard feedforward network with ReLU activations. Denoting the MLP as $f_\phi$:

$$MLP(O) = f_\phi\left(\frac{1}{N}\sum_{i=1}^{N} e_\omega(o_i)\right)$$

*Convolutional Neural Network (CNN):* this model encodes a translation invariance bias, receiving the objects as a 2D grid $S$ rather than as an unordered list of vectors. As the objects' positions are represented by their placement in the grid, we use two channels in the spatial input, one marking the target object's location and another marking the locations of all reference objects. We use a standard convolutional architecture (conv) followed by global average pooling and an MLP:

$$CNN(S) = f_\phi(\text{average\_pool}(\text{conv}(S)))$$

*Relation Net:* Santoro et al. (2017) offer a compact way of modeling relations between pairs of objects, using two functions: a function $g_\theta$ that acts on object pairs and a global MLP $f_\phi$ acting on their combined representation:

$$RN(O) = f_\phi\left(\sum_{i=1}^{N}\sum_{j=1}^{N} g_\theta\left(e_\omega(o_i), e_\omega(o_j)\right)\right)$$

*Transformer:* a simplification of the Transformer (see Vaswani et al., 2017 for details), this network reasons about all objects jointly rather than through object pairs. The self-attention ('SelfAttn' below) operator acts on the entire set of objects simultaneously to capture their interactions. We pass the input through one or more such Encoders, and the transformed representations are averaged and passed through a MLP:

$$\text{Encoder}(E) = E + \text{SelfAttn}(E) + f_\phi\left(E + \text{SelfAttn}(E)\right)$$

$$T(O) = f_\phi\left(\frac{1}{N}\sum_{i=1}^{N} \overbrace{\text{Encoder}(e_\omega(O))}^{\text{One or more times}}\right)$$

*PrediNet:* this model is explicitly designed to learn different relations between objects, making for a task-optimized comparison architecture. It uses a modified form of self-attention, combining global information over the entire set of objects with information from each individual object, and treats the difference between object representations in a latent space as capturing different relations between them. See Shanahan et al. (2019) for the full details.

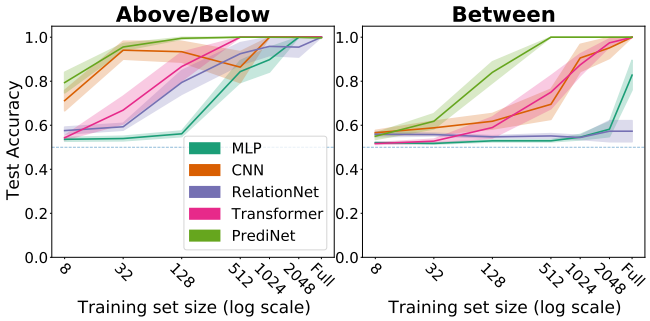


**Figure 3: Smaller model test set accuracy by training set size.** Left: *above/below*. Right: *between*. Average over ten random seeds, shaded regions mark the SEM. Dashed line indicates chance (50%).

**Implementation and Training.** To test the effect of model size, we created two configurations of each model, a smaller one (using around 2000 parameters) and a larger one (using around 8000 parameters). We report results from ten random seeds for each simulation, varying three factors: *above/below* or *between*, smaller or larger models, and the number of training examples. We terminated each run when performance on the validation set plateaued. All models were optimized using Adam Kingma and Ba (2015) with a learning rate of 1e-3 and a batch size of 256. All models were implemented in PyTorch (Paszke et al., 2017) using PyTorch Lightning (Falcon, 2019).

## Results

We focus our analysis on two measures of learning difficulty: *Sample complexity:* how many examples does it take to learn each concept? *Number of epochs:* how many passes through the training set does it take to learn each concept? We evaluate all five models on their ability to capture the developmental phenomena described above, including which architectures may be too powerful (learning both conditions trivially) or too weak (failing to learn either condition) when compared with competencies in infancy.

To evaluate the sample complexity, Figure 3 depicts the test set accuracy attained by each architecture as a function of the size of the training set used, using the smaller (2000 parameter) configurations. We plot only the test set accuracies, as the networks generalize well above a reasonable sample size: the maximal difference between the training and test accuracy, averaged over the replications of each network, is 12.7% with 128 samples, 2.4% with 512 samples, and $< 0.1\%$ with the full training set. At all training set sizes, the networks perform better in the *above/below* condition than they do in the *between* condition, unless they fail to learn both. This is true from the most successful network (PrediNet) to the simplest (MLP) one, using both the smaller and larger network configurations. The RelationNet is the only network which fails to learn a relation, never reaching much above chance accuracy in the *between* condition; the MLP also struggles with *between*, rising above chance only with the full dataset. Results using the larger model configurations showed the same qualitative patterns.

To explore how long it takes the networks to acquire the

concepts, Figure 4 illustrates the learning curves using a 1024-item training set. Unsurprisingly, the models that reach a higher test accuracy (Figure 3) also tend to require fewer training epochs to reach high performance. All architectures reach peak accuracy faster in the *above/below* condition than in the *between* condition. At this dataset size, both the RelationNet and the MLP networks fail to learn in the *between* condition.

### Discussion

Most of the architectures examined are consistent with the basic developmental phenomenon: learning to spatially categorize above versus below is easier than between versus outside. This holds both when we take the sample complexity as a proxy for experience, and when we take the number of training epochs as the measure of experience. The RelationNet model struggled with learning the *between* relation, suggesting it may be an inadequate model of infant relation learning. In our alternative object formulation (see "Objects" in the Methodology subsection), which adds a length entry and hence reduces the number of input vectors, the RelationNet succeeded to learn this relation, performing closer to the Transformer. We therefore attribute this failure to the fact that learning to reason using a pairwise function over the objects is harder to scale to higher numbers of entities. Models that natively reason over the entire collection struggle less with the *between* relation, which requires comparing three objects, the "local spatial framework" discussed by Huttenlocher and Newcombe (1984). The CNN and the Transformer both recover patterns qualitatively resembling the developmental findings, as does the PrediNet, even though it requires substantially less data than the other architectures to reach perfect accuracy. Conversely, the MLP might be overly generic, as it struggles with the *between* condition, only reaching above-chance performance with the full training set. We take these results to imply that any compelling computational model of infant reasoning should flexibly allow for variation in the number of objects reasoned over, being neither entirely generic (the MLP) nor restricted to pairwise interactions (the RelationNet). Beyond these constraints and considerations, the data does not help us distinguish the other architectures as potential cognitive models. The finding that learning above/below is easier than between/outside appears to be a fairly general property of the neural architectures we evaluated.

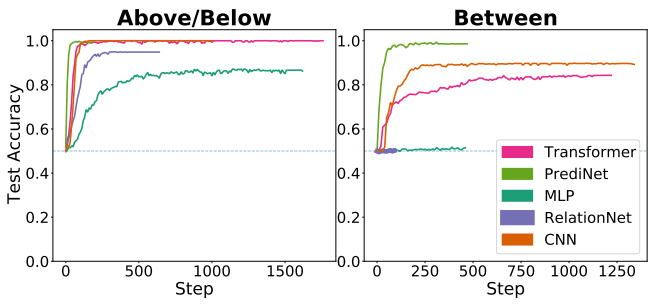


**Figure 4: Smaller model learning curves using 1024 training items.** Left: *above/below*. Right: *between*. Average over ten random seeds for each model. Dashed line indicates chance accuracy (50%).

## Experiment 2: Pretrained Vision Models

In this experiment, in addition to the finding modeled in experiment 1, we also model the second finding discussed, that infants can encode relations for specific objects earlier than they can encode abstract spatial categories over different objects. Quinn et al. (1996) extended the above/below relation experiment of Quinn (1994), using different target objects in each familiarization and test example. As before, the two test stimuli displayed a target object that either matched or differed from the habituation stimuli in terms of spatial relation. Infants 3-4 months old did not display a preference to either of the new stimuli; infants 6-7 months old demonstrated a significant preference of the novel spatial category. Quinn et al. (2003) performed a similarly modified version of the between experiment reported by Quinn et al. (1999), and found similar results. Whereas 6-7 month old infants did not appear to construct a spatial category abstract of the particular stimulus, 9-10 months old infants reproduce the novelty preference to the stimuli depicting the target outside the references.

### Methodology

We examine these developmental phenomena through a different class of models: large pretrained convolutional neural networks. In the previous experiment, we trained small models directly on learning to classify relations, as a proxy for infants learning spatial categories outside the lab over an extended number of examples. In this experiment, we use pretrained models to examine which phenomena could arise absent explicit training on relations; instead the ability to discriminate different relations in the lab could emerge from high-level visual representations developed for other purposes. In each triplet (Figure 5), one image (left column) corresponds to a familiarization example while the other two are test probes: one (middle column) presents the same relation with the target object in a different location, and the other (right column) presents a different relation. During evaluation, we do not explicitly task the models with predicting which relation holds in each image. Instead, we extract latent representations of the stimuli to see if stimuli with the same relation are represented more similarly, as an emergent consequence of training a network on broad visual experience (in one case, of the sort one baby would actually experience). Specifically, we take the cosine similarity between vector embeddings of the familiarization example and each of the other two images. We consider a triplet to be accurately classified when the model embeds the two congruent images (depicting the same relation) more similarly than the two incongruent images (depicting different relations).

**Architectures.** We evaluate two computer vision architectures: *MobileNetV2:* this model aims to offer competitive performance with fairly limited computational resources, offering state of the art trade-off between compute resources required and performance attained (Sandler et al., 2018). *ResNeXt:* this architecture does not strive for performance over efficiency and is considered one of the best-performing vision backbones

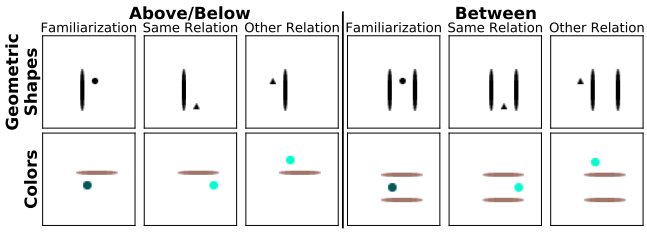


**Figure 5: Experiment 2 Stimuli.** Similar in structure to Figure 1. Left columns: *above/below*. Right columns: *between*. Top row: *Geometric shapes*. Bottom row: *Colors*. For the *Quinn-like* rendering, see Figure 1; bottom row. We visualize vertical examples (left/right) in the first row and horizontal ones (above/below) in the second row.

(Xie et al., 2017).

**Pretraining Datasets.** As a baseline, we test the embeddings created by randomly initialized models, examining whether or not the inductive biases conveyed by the architecture are sufficient to embed objects in the same relation more similarly. We then then compare these results to models trained on the following datasets: *SAYCam:* this dataset offers longitudinal headcam video from a small number of babies (Sullivan et al., 2020). We use models trained on a single child's footage (child S), approximately two hours per week while the child was between 6-30 months old, a total of 221 hours. This offers the opportunity to train vision models on a subset of the experience a child receives in development, albeit ranging to older ages than the infants studied in the experiments modeled. We utilize a pretrained network from Orhan et al. (2020) that uses temporal classification, a self-supervised approach that requires no category labels. *ImageNet:* a landmark computer vision dataset, offering 1.2M images in 1000 object classes (Russakovsky et al., 2015). ImageNet does not resemble an infant's experience in development, but it is often used for general computer vision pretraining, offering a useful comparison. The ImageNet models were pretrained using the standard classification task as described in the torchvision documentation.

**Stimulus Generation.** We synthesize custom stimuli to probe the model in this task (Figure 5). As in experiment 1, we sample locations for the reference object(s) and then place the target objects relative to them. Similarly to Quinn (1994; 1996; 1999), we place the target object in one relation relative to the reference object in the familiarization example, and then place it in a different location in the same relation (first test probe) or in the other relation (second test probe). The target objects in the test probes are both equidistant from the target object in the familiarization probe, controlling for any effect of distance on the representational similarity. We examine triplets where the target object matches between the familiarization and probe stimuli ("same target"; Quinn, 1994; Figure 1; top) and triplets where the probe stimuli use a different target object ("different target"; Quinn et al., 1996; Figure 1; bottom).

We explore a few ways to render the reference and target objects: *Quinn-like:* Most similar to Quinn et al. (1996), we render the reference object as a sequence of squares and the target object as one of the symbols used in that paper (a triangle, 's', 'E', +, and →), all colored black (Figure 1; bottom row). *Geometric shapes:* we render the reference as an elongated ellipse and the target as either a square, a circle, or a triangle, all colored black (Figure 5; top row). *Colors*: again we render the reference as an elongated ellipse and the target as a circle, sampling perceptually distinct colors for both using the glasbey method (Glasbey et al., 2007; Kovesi, 2015, Figure 5, bottom row). The latter deviates most from the original formulations, but allows programmatically sampling a larger variety of stimuli to verify result robustness. We experimented with slightly blurring the stimuli to make them less perceptually perfect; this did not substantially impact any results. We render these stimuli to 224x224 pixel images.

To summarize, we evaluate models from both architectures, either randomly initialized or pretrained on one of the visual datasets, in two horizontal conditions (*above/below* and *between*) and two vertical conditions (*left/right* and *vertical-between*). In each condition and object rendering method, we sample 1024 triplets and report the average accuracy for each model and pretraining fashion—how often are the embeddings for the congruent pair of stimuli are more similar (using cosine similarity) than the embeddings for the incongruent pair. With the *colors* rendering, we repeat this 10 times with different colors.

## Results

Without pretraining, the models perform effectively at chance—the ResNeXts range between 40.5% to 66.5%, and the MobileNetV2s between 37.8% to 53.3%, suggesting that any inductive bias conferred by the architecture alone in insufficient for this task. We therefore focus on the ImageNet and SAYCam-pretrained models. Figure 6 summarizes the results in the two conditions with horizontal reference objects, *above/below* and *between*. As observed patterns were similar between the three rendering methods used, we average over them in both result figures. We find a higher accuracy in each *above/below* group than in the corresponding *between* group, with differences ranging from 0.9% to 15%. We also observe higher accuracies with an identical target object between the familiarization and the probe stimuli (without hatches) than in the conditions with different target objects (with hatches). These two findings are consistent with the two main developmental results discussed. Lastly, we see that SAYCam pretraining results in higher accuracies than ImageNet pretraining, and that the MobileNetV2 models outperform the ResNeXt models.

Figure 7 similarly summarizes results in the conditions with vertical references, *left/right* and *vertical-between*. Surprisingly, the accuracy levels are much worse overall, with most conditions either near or below chance. The only pattern that holds from the previous results is the advantage of the MobileNetV2 models over the ResNeXt ones. Other findings do not replicate—the models reach higher accuracy in the *vertical-between* condition than in the *left/right* condition, and we mostly see slightly higher accuracies when using different target objects than with the same target objects.

## Discussion

At first, we found the results in the horizontal conditions quite compelling: both main developmental findings we model replicated clearly, across two different models and pretraining datasets. We then found the discrepancy between the vertical conditions and the horizontal conditions, which is inconsistent with the developmental findings modeled. Quinn et al. (1999) found similar results with vertical and horizontal references, and Quinn (2004) replicated the findings of Quinn et al. (1996) using left versus right instead of above versus below, although Landau and Hoffman (2005) offer evidence that left versus right can be harder than above versus below. One hypothesis that could account for some of this deviation arises from the use of data augmentation. The standard suite of data augmentation transformations for computer vision, used during pretraining for the ResNeXt models (but not the MobileNet models) by both Orhan et al. (2020) and the torchvision library, includes horizontal flipping but not vertical flipping. This influences the models to learn representations invariant to horizontal flipping (as an image and its horizontally flipped counterpart are assigned the same class), without instilling a similar bias toward vertical flipping, potentially accounting for the difference between the conditions. We find it compelling that the SAYCam models, trained from the perspective of a single child, outperform the models trained on over a million ImageNet images. Although the difference could arise from the naturalistic, egocentric perspective of images in the SAYCam dataset, it could also result from the different procedures used. Orhan et al. (2020) employed a self-supervised temporal classification procedure for training on SAYCam, compared to the supervised learning procedure for training on ImageNet.

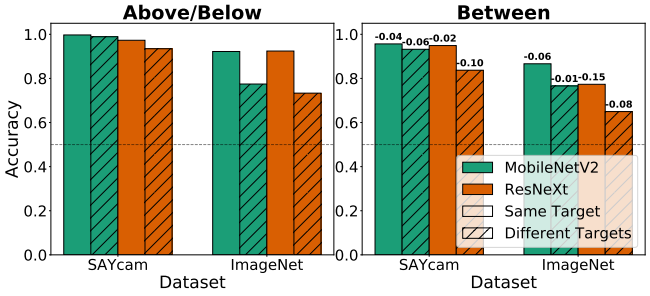


**Figure 6: Horizontal references.** Left: *above/below*. Right: *between*. Bar groups: pretraining dataset. Color: model architecture. Hatching: same or different targets. Numbers in the *between* panel indicate accuracy differences from the *above/below* condition.

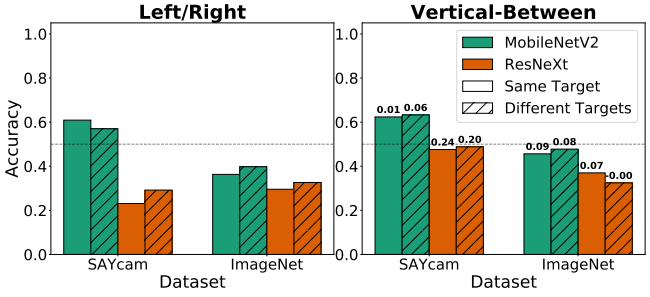


**Figure 7: Vertical references.** Identical to Figure 6, but for the *left/right* and *vertical-between* conditions.

## General Discussion

We report results from two computational experiments modeling infant relation learning. In experiment 1, we train small-scale deep neural networks from scratch, as a proxy for acquiring more explicit representations for relations outside the lab. In experiment 2, we probe large-scale, pretrained computer vision models as a proxy for using features acquired for other purposes to quickly construct a category representation during familiarization trials. As both experiments are consistent with the key developmental findings surveyed, we view this as evidence that a range of different neural network models can account for these findings, regardless of whether infants acquire their knowledge of relations before or during the lab session.

In future work, we hope to extend the from-scratch paradigm from experiment 1 to engage with the sorts of abstract relations studied in experiment 2, perhaps using a meta-learning setup, to observe if models trained from scratch can account for these additional findings. In this proposed setup, we would examine whether models that learn to classify a relation with a variety of target objects, can learn to classify novel objects in the same relation in a few-shot or zero-shot fashion. We also plan to investigate the discrepancy between categorization relations with horizontal stimuli (*above/below* and *between*) versus relations with vertical stimuli (*left/right* and *vertical-between*) in experiment 2, by examining the roles of vertical and horizontal flipping used in data augmentation. We will also examine using multiple familiarization examples, represented by the mean of their embeddings or through another method of aggregation, to more closely imitate the developmental experiments. Additionally, the role of the pretraining dataset (ImageNet or SAYCam) is currently confounded with the choice of pretraining procedure used (supervised learning or temporal classification), which we hope to dissociate. Finally, recent work identified modifications to convolutional neural network architectures that better model the visual stream (Kar et al., 2019; Dapello et al., 2020). It would be interesting to compare more biologically-driven models to the more standard architectures we used, examining whether neural plausibility associates with the ability to capture behavioral findings.

## Acknowledgements

The authors would like to thank Agata Bochynska, Reuben Feinman, Emin Orhan, and Wai Keen Vong for helpful feedback on earlier versions of this manuscript. This work was supported by the DARPA Machine Common Sense program and NSF Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science.

## References

Barrett, D. G. T., Hill, F., Santoro, A., Morcos, A. S., and Lillicrap, T. (2018). "Measuring abstract reasoning in neural networks". *ICML 2018* 10 (p. 1).

Battaglia, P. W., Hamrick, J. B., Bapst, V., et al. (2018). "Relational inductive biases, deep learning, and graph networks" (p. 1).

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D. D., and DiCarlo, J. J. (2020). "Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations". *bioRxiv*, pp. 1–26 (p. 6).

Falcon, W. (2019). "PyTorch Lightning". *GitHub. Note: https://github.com/williamFalcon/pytorch-lightning Cited by* 3 (p. 3).

French, R. M., Mermillod, M., Mareschal, D., and Quinn, P. C. (2004). "The role of bottom-up processing in perceptual categorization by 3- to 4-month-old infants: Simulations and data". *Journal of Experimental Psychology: General* 133.3, pp. 382–397 (p. 1).

Glasbey, C., Heijden, G. van der, Toh, V. F. K., and Gray, A. (2007). "Colour displays for categorical images". *Color Research & Application* 32.4, pp. 304–309 (p. 5).

Goodwin, G. P. and Johnson-Laird, P. N. (2005). "Reasoning about relations." *Psychological Review* 112.2, pp. 468–493 (p. 1).

Hamrick, J. B., Allen, K. R., Bapst, V., Zhu, T., McKee, K. R., Tenenbaum, J. B., and Battaglia, P. W. (2018). "Relational inductive bias for physical construction in humans and machines" (p. 1).

Huttenlocher, J. and Newcombe, N. S. (1984). "The child's representation of information about location". In: *Origin of cognitive skills*. Ed. by C. Sophian. Hillsdale, NJ: Erlbaum, pp. 81–111 (pp. 2, 4).

Johnson, S. P. (2010). "How Infants Learn About the Visual World". *Cognitive Science* 34.7, pp. 1158–1184 (p. 1).

Kar, K., Kubilius, J., Schmidt, K. M., Issa, E. B., and DiCarlo, J. J. (2019). "Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior". *Nature Neuroscience* 22, pp. 974–983 (p. 6).

Kingma, D. P. and Ba, J. (2015). "Adam: A Method for Stochastic Optimization". In: *ICLR 2015* (p. 3).

Kipf, T., Fetaya, E., Wang, K.-C., Welling, M., and Zemel, R. (2018). "Neural Relational Inference for Interacting Systems". *ICML 2018* (p. 1).

Kovesi, P. (2015). "Good Colour Maps: How to Design Them" (p. 5).

Landau, B. and Hoffman, J. E. (2005). "Parallels between spatial cognition and spatial language: Evidence from Williams syndrome". *Journal of Memory and Language* 53.2, pp. 163–185 (p. 6).

Mareschal, D., French, R. M., and Quinn, P. C. (2000). "A connectionist account of asymmetric category learning in early infancy." *Developmental psychology* 36.5, pp. 635–645 (p. 1).

Murphy, G. (2002). *The Big Book of Concepts*. Cambridge, MA: MIT Press (p. 1).

Newcombe, N. S. and Huttenlocher, J. (2007). "Development of Spatial Cognition". In: *Handbook of Child Psychology*. Hoboken, NJ, USA: John Wiley & Sons, Inc. (p. 1).

Newcombe, N. S., Sluzenski, J., and Huttenlocher, J. (2005). "Preexisting knowledge versus on-line learning: What do young infants really know about spatial location?" *Psychological Science* 16.3, pp. 222–227 (p. 1).

Orhan, A. E., Gupta, V. V., and Lake, B. M. (2020). "Self-supervised learning through the eyes of a child". In: *NeurIPS 2020*. arXiv (pp. 5, 6).

Paszke, A., Gross, S., Chintala, S., et al. (2017). "Automatic differentiation in PyTorch". In: *NeurIPS 2017* (p. 3).

Piaget, J. (1954). *The construction of reality in the child*. New York, NY: Basic Books (p. 1).

Quinn, P. C. (1994). "The Categorization of Above and Below Spatial Relations by Young Infants". *Chil Dev* 65.1, pp. 58–69 (pp. 1, 2, 4, 5).

– (2003). "Concepts are not just for objects: Categorization of spatial relation information by infants". In: *Early category and concept development: Making sense of the blooming, buzzing confusion*. Ed. by D. H. Rakison and L. M. Oakes. Oxford University Press. (p. 1).

Quinn, P. C. (2004). "Spatial representation by young infants: Categorization of spatial relations or sensitivity to a crossing primitive?" *Memory and Cognition* 32.5, pp. 852–861 (pp. 1, 6).

Quinn, P. C., Adams, A., Kennedy, E., Shettler, L., and Wasnik, A. (2003). "Development of an abstract category representation for the spatial relation between in 6- to 10-month-old infants." *Developmental Psychology* 39.1, pp. 151–163 (pp. 1, 4).

Quinn, P. C., Cummins, M., Kase, J., Erin, M., and Weissman, S. (1996). "Development of categorical representations for above and below spatial relations in 3- to 7-month-old infants". *Developmental Psychology* 32.5, pp. 942–950 (pp. 1, 2, 4–6).

Quinn, P. C., Norris, C. M., Pasko, R. N., Schmader, T. M., and Mash, C. (1999). "Formation of a categorical representation for the spatial relation between by 6- to 7-month-old infants". *Visual Cognition* 6.5, pp. 569–585 (pp. 2, 4–6).

Regier, T. (1995). "A model of the human capacity for categorizing spatial relations". *Cognitive Linguistics* 6.1, pp. 63–88 (p. 1).

Russakovsky, O., Deng, J., Su, H., et al. (2015). "ImageNet Large Scale Visual Recognition Challenge". In: *IJCV* (p. 5).

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). "MobileNetV2: Inverted Residuals and Linear Bottlenecks". *CVPR 2018* (p. 4).

Santoro, A., Raposo, D., Barrett, D. G. T., et al. (2017). "A simple neural network module for relational reasoning". In: *NeurIPS 2017* (pp. 1, 3).

Shanahan, M., Nikiforou, K., Creswell, A., Kaplanis, C., Barrett, D., and Garnelo, M. (2019). "An Explicitly Relational Neural Network Architecture" (pp. 1, 3).

Sullivan, J., Mei, M., Perfors, A., Wojcik, E., and Frank, M. (2020). "SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective" (pp. 1, 5).

Teodorescu, L., Hofmann, K., and Oudeyer, P.-Y. (2020). "Recognizing Spatial Configurations of Objects with Graph Neural Networks" (p. 1).

Thelen, E. and Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press (p. 1).

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). "Attention Is All You Need". In: *NeurIPS 2017*. Long Beach, CA, USA (p. 3).

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). "Aggregated residual transformations for deep neural networks". In: *CVPR 2017*, pp. 5987–5995 (p. 5).

Yuan, L., Uttal, D., and Gentner, D. (2017). "Analogical processes in children's understanding of spatial representations". *Developmental Psychology* 53.6, pp. 1098–1114 (p. 1).