Author for correspondence:
Thomas Serre
e-mail: thomas_serre@brown.edu

†These authors contributed equally to this study.

Not-So-CLEVR: learning same–different relations strains feedforward neural networks

Junkyung Kim†, Matthew Ricci† and Thomas Serre

Department of Cognitive, Linguistic & Psychological Sciences, Carney Institute for Brain Science, Brown University, Providence, RI 02912, USA

TS, 0000-0003-0846-0039

The advent of deep learning has recently led to great successes in various engineering applications. As a prime example, convolutional neural networks, a type of feedforward neural network, now approach human accuracy on visual recognition tasks like image classification and face recognition. However, here we will show that feedforward neural networks struggle to learn abstract visual relations that are effortlessly recognized by non-human primates, birds, rodents and even insects. We systematically study the ability of feedforward neural networks to learn to recognize a variety of visual relations and demonstrate that same–different visual relations pose a particular strain on these networks. Networks fail to learn same–different visual relations when stimulus variability makes rote memorization difficult. Further, we show that learning same–different problems becomes trivial for a feedforward network that is fed with perceptually grouped stimuli. This demonstration and the comparative success of biological vision in learning visual relations suggests that feedback mechanisms such as attention, working memory and perceptual grouping may be the key components underlying human-level abstract visual reasoning.

1. Introduction

Consider the images in figure 1*a*. These images were correctly classified as two different breeds of dog by a state-of-the-art computer vision system called a convolutional neural network (CNN) [3]. This is quite a remarkable feat because the network must learn to extract subtle diagnostic cues from images subject to a wide variety of factors such as scale, pose and lighting. The network was trained on millions of photographs, and images such as these were accurately categorized into 1000 natural object labels, surpassing, for the first time, the accuracy of a human observer for the recognition of 1000 image categories on the ImageNet classification challenge [4].

Now, consider the image in figure 1*b*(i). On its face, it is quite simple compared with the images in figure 1*a*. It is just a binary image containing two three-dimensional shapes. Further, it has a rather distinguishing property: both shapes are the same up to rotation. The relation between the two items in this simple scene is rather intuitive and obvious to human and non-human observers. In a recent, striking example from Martinho & Kacelnik [2], newborn ducklings were shown to imprint on an abstract concept of 'sameness' from a single training example at birth (figure 1*b*(ii)). Yet, as we will show in this study, CNNs struggle to learn this seemingly simple concept.

Why is it that a CNN can accurately categorize natural images while struggling to recognize a simple abstract relation? That such a task is difficult or even impossible for contemporary computer vision algorithms is known. Previous work by Fleuret *et al.* [5] has shown that black-box classifiers fail on most tasks from the Synthetic Visual Reasoning Test (SVRT), a battery of 23 visual-relation problems, despite massive amounts of training data. More
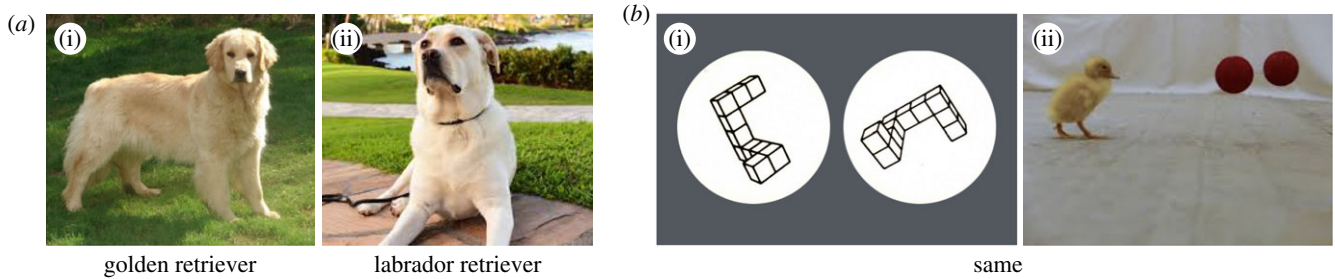
Figure 1. (*a*) State-of-the-art convolutional neural networks can learn to categorize images (including dog breeds) with high accuracy even when the task requires detecting subtle visual cues. The same networks struggle to learn the visual recognition problems shown in (*b*). (*b*) In addition to categorizing visual objects, humans can also perform comparison between objects and determine if they are identical up to a rotation (i). The ability to recognize 'sameness' is also observed in other species in the animal kingdom such as birds (ii). (*b*(i)) Adapted from [1]; (*b*(ii)) taken with permission from Martinho & Kacelnik [2].

recent work has shown how CNNs, including variants of the popular LeNet [6] and AlexNet [7] architectures, could only solve a handful of the 23 SVRT problems [8,9]. Similarly, Gülçehre & Bengio [10], after showing how CNNs fail to learn a same−different task with simple binary 'sprite' items, only managed to train a multi-layer perceptron on this task by providing carefully engineered training schedules.

However, these results are not entirely conclusive. First, each of these studies only tested a small number of feedforward architectures, leaving open the possibility that low accuracy on some of the problems might simply be a result of a poor choice of model hyper-parameters. Second, while the 23 SVRT problems represent a diverse collection of relational concepts, the images used in each problem are also visually distinct (e.g. some relations require stimuli to have three items, while others require two). This makes a direct comparison among different problems challenging because the performance of a computational model on a given problem may be driven by specific features in that problem rather than the underlying abstract rule. To our knowledge, there has been no systematic exploration of the limits of contemporary machine learning algorithms on relational reasoning problems. Additionally, the issue has been overshadowed by the recent success of novel architectures called relational networks (RNs) on seemingly challenging visual question answering benchmarks [11].

In this study,[1] we probe the limits of feedforward neural networks, including CNNs and RNs, on visual-relation tasks. In experiment 1, we perform a systematic performance analysis of CNN architectures on each of the 23 SVRT problems, which reveals a dichotomy of visual-relation problems: hard same−different problems and easy spatial-relation problems. In experiment 2, we introduce a novel, controlled, visual-relation challenge called parametric SVRT (PSVRT), which we use to demonstrate that CNNs solve same−different tasks only inefficiently, via rote memorization of all possible spatial arrangements of individual items. In experiment 3, we examine two models, the RN and a novel Siamese network, which simulate the effects of perceptual grouping and attentional routing to solve visual-relation problems. We find that the former struggles to learn the notion of sameness and tends to overfit to particular item features, but that the latter can render seemingly difficult visual reasoning problems rather trivial.

Overall, our study suggests that a critical reappraisal of the capability of current machine vision systems is warranted. We further argue that mechanisms for individuating objects and manipulating their representations, presumably through feedback processes that are absent in current feedforward architectures, are necessary for abstract visual reasoning.

2. Experiment 1: A dichotomy of visual-relation problems

2.1. The SVRT challenge

The SVRT is a collection of 23 binary classification problems in which opposing classes differ based on whether or not images obey an abstract rule [5]. For example, in problem number 1, positive examples feature two items which are the same up to translation (figure 2), whereas negative examples do not. In problem 9, positive examples have three items, the largest of which is in between the two smaller ones. All stimuli depict simple, closed, black curves on a white background.

For each of the 23 problems, we generated 2 million examples split evenly into training and test sets using code made publicly available by the authors of the original study at http://www.idiap.ch/fleuret/svrt.

2.2. Hyper-parameter search

We tested nine different CNNs of three different depths (two, four and six convolutional layers) and with three different convolutional filter sizes (2×2, 4×4 and 6×6) in the first layer. This initial receptive field size effectively determines the size of receptive fields throughout the network. The number of filters in the first layer was six, 12 or 18, respectively, for each choice of initial receptive field size. In the other convolutional layers, filter size was fixed at 2×2 with the number of filters doubling every layer. All convolutional layers had strides of 1 and used rectified linear (ReLU) activations. Pooling layers were placed after every convolutional layer, with pooling kernels of size 3×3 and strides of 2. On top of the retinotopic layers, all nine CNNs had three fully connected layers with 1024 hidden units in each layer, followed by a two-dimensional classification layer. All CNNs were trained on all problems. Network parameters were initialized using Xavier initialization [13] and were trained using the Adaptive Moment Estimation (Adam) optimizer [14] with a base learning rate of $\eta = 10^{-4}$. All experiments were run using TensorFlow [15].

2.3. Results

Figure 3 shows a ranked bar plot of the best-performing network accuracy for each of the 23 SVRT problems. Bars
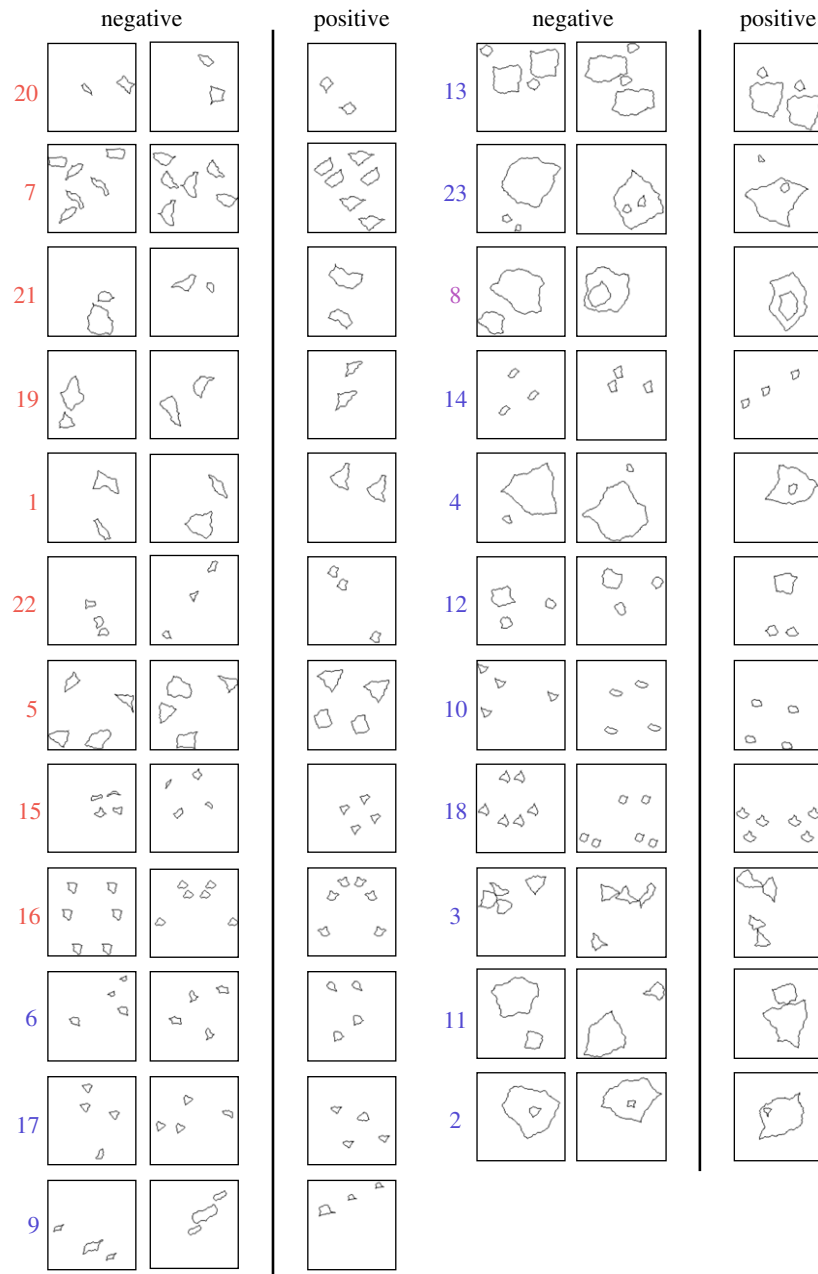
Figure 2. Sample images from the 23 SVRT problems. For each problem, three example images, two negative and one positive, are displayed in a row. Problems are ordered and colour-coded identically to figure 3. Images in each problem all respect a certain visual structure (e.g. in problem 9, three objects, identical up to a scale, are arranged in a row). Positive and negative categories are then characterized by whether or not objects in an image obey a rule (e.g. in problem 3, an image is considered positive if it contains two touching objects and negative if it contains three touching objects). Descriptions of all problems can be found in [5].

are coloured red or blue according to the SVRT problem descriptions given in [5]. Problems whose descriptions have words like 'same' or 'identical' are coloured red. These *same–different* (SD) problems have items that are congruent up to some transformation. *Spatial-relation* (SR) problems, whose descriptions have phrases such as 'left of', 'next to' or 'touching', are coloured blue. Figure 2 shows positive and negative samples for each of the 23 problems (also sorted by network accuracy from low to high).

The resulting dichotomy across the SVRT problems is striking. CNNs fare uniformly worse on SD problems than they do on SR problems. Many SR problems were learned satisfactorily, whereas some SD problems (e.g. problems 20, 7) resulted in accuracy not substantially above chance. From this analysis, it appears as if SD tasks pose a particularly difficult challenge to CNNs. This is consistent with results from an earlier study by Stabinger *et al.* [9].

Additionally, our search revealed that SR problems are equally well learned across all network configurations, with less than 10% difference in final accuracy between the worst and the best network. On the other hand, deeper networks yielded significantly higher accuracy on SD problems than on smaller ones, suggesting that SD problems require a higher capacity than SR problems. Experiment 1 corroborates the results of previous studies which found feedforward neural networks performed badly on many visual-relation problems [5,8–11] and suggests that low accuracy cannot be simply attributed to a poor choice of hyper-parameters.

2.4. Limitations of the SVRT challenge

Although useful for surveying many types of relations, the SVRT challenge has two important limitations. First, different problems have different visual structures. For instance,
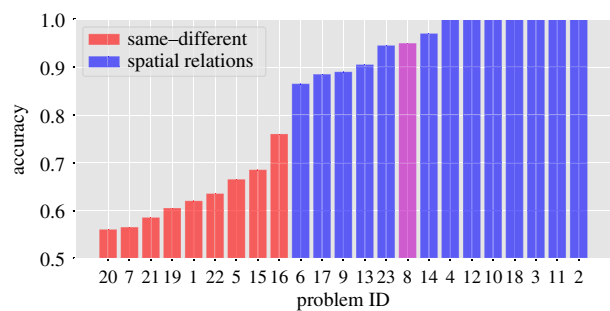
Figure 3. SVRT results. Multiple CNNs with different combinations of hyper-parameters were trained on each of the 23 SVRT problems. Shown are the ranked accuracies of the best-performing networks optimized for each problem individually. The *x*-axis shows the problem ID. CNNs from this analysis were found to produce uniformly lower accuracies on same–different problems (red bars) than on spatial-relation problems (blue bars). The purple bar represents a problem which required detecting both a same–different relation and a spatial relation.

problem 2 (*inside–outside*) requires that an image contain one large item and one small item. Problem 1 (*same–different up to translation*), on the other hand, requires that an image contain two items, identically sized and positioned without one being contained in the other. In other cases, different problems simply require a different number of items in a single image (two items in problem 1 versus three in problem 9). This confound leaves open the possibility that image features, not abstract relational rules, make some problems harder than others. Instead, a better way to compare visual-relation problems would be to define various problems on the *same* set of images. Second, the ad hoc procedure used to generate simple, closed curves as items in SVRT prevents quantification of image variability and its effect on task difficulty. As a result, even within a single problem in SVRT, it is unclear whether its difficulty is inherent to the classification rule itself or simply results from the particular choice of image generation parameters unrelated to the rule.

3. Experiment 2: A systematic comparison between spatial-relation and same–different problems

3.1. The PSVRT challenge

To address the limitations of SVRT, we constructed a new visual-relation benchmark consisting of two idealized problems (figure 4) from the dichotomy that emerged from experiment 1: SR and SD. Critically, both problems used exactly the same images, but with different labels. Further, we parameterized the dataset so that we could systematically control various image parameters; namely, the size of scene items, the number of scene items and the size of the whole image. Items were binary bit patterns placed on a blank background.

For each configuration of image parameters, we trained a new instance of a single CNN architecture and measured the ease with which it fit the data. Our goal was to examine how hard it is for a CNN architecture to learn relations for visually different but conceptually equivalent problems. For example, imagine two instances of the same CNN architecture, one trained on a same–different problem with small items in a

large image, and the other trained on large items in a small image. If the CNNs can truly learn the 'rule' underlying these problems, then one would expect the models to learn both problems with more or less equal ease. However, if the CNNs only memorize the distinguishing features of the two image classes, then learning should be affected by the variability of the example images in each category. For example, when image size and item size are large, there are simply more possible samples, which might put a strain on the representational capacity of a CNN trying to learn by rote memorization.

In rule-based problems such as visual relations, these two strategies can be distinguished by training and testing the same architecture on a problem instantiated over a multitude of image distributions. Here, our main question is not whether a model trained on one set of images can accurately predict the labels of another, unseen set of images sampled from the same distribution. Rather, we want to understand whether an architecture that can easily learn a visual relation instantiated from one image distribution (defined by one set of image parameters) can also learn the same relation instantiated from another distribution (defined by another set of parameters) with equal ease by taking advantage of the abstractness of the visual rule. Evidence that CNNs use rote memorization of examples was found in a study by Stabinger & Rodriguez-Sanchez [16], who tested state-of-the-art CNNs on a type of same–different problem using a dataset of realistically rendered images of checkerboards. Stabinger & Rodriguez-Sanchez [16] found that CNN accuracy was lower on datasets whose images were rendered with higher degrees of freedom in viewpoint. In our study, we take a similar approach while using much simpler synthetic images where we can explicitly compute intra-class variability as a function of image parameters. This way, we do not introduce any additional perceptual nuisances such as specularity or three-dimensional rotation whose contribution to image variability and CNN performance is difficult to quantify. Because PSVRT images are randomly synthesized, we generate training images online without explicitly reusing data, and there is no hold-out set in this experiment. Thus, we use training accuracy to measure the ease with which a model learns a visual-relation problem.

3.2. Methods

Our image generator produces a grey-scale image by randomly placing square binary bit patterns (consisting of values 1 and −1) on a blank background (with value 0). The generator uses three parameters to control image variability: the size (m) of each bit pattern or item, the size (n) of the input image and the number (k) of items in an image. Our parametric construction allows a dissociation between two possible factors that may affect problem difficulty: classification rules versus image variability. To highlight the parametric nature of the images, we call this new challenge the *parametric SVRT* or *PSVRT*.

Additionally, our image generator is designed such that each image can be used to pose both problems by simply labelling it according to different rules (figure 4). In SR, an image is classified according to whether the items in an image are arranged horizontally or vertically as measured by the orientation of the line joining their centres (with a 45° threshold). In SD, an image is classified according to whether or not it contains at least two identical items. When $k \geq 3$, the SD category label is determined by whether or not there are *at least* two identical items in the image, and the SR category label is determined
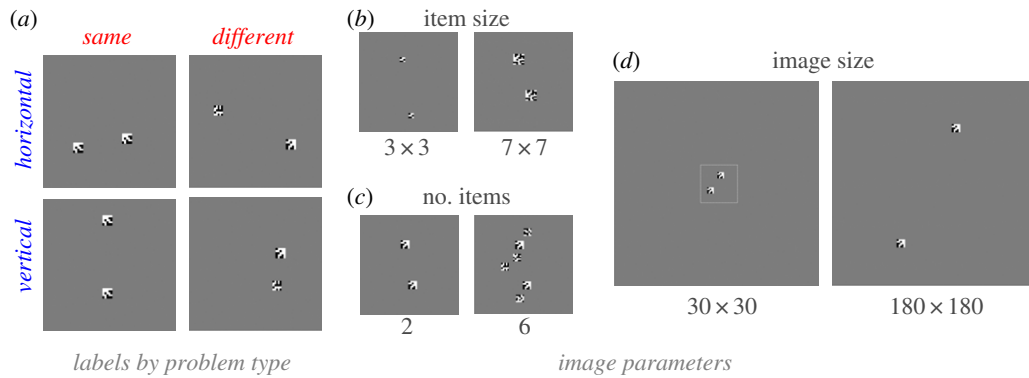
Figure 4. The PSVRT challenge. (*a*) Four images show the joint categories of SD (grouped by columns) and SR (grouped by rows) tasks. Our image generator is designed so that each image can be used to pose both problems by simply labelling it according to different rules. An image is *same* or *different* depending on whether it contains identical (left column) or different (right column) square bit patterns. An image is *horizontal* (top row) or *vertical* (bottom row) depending on whether the orientation of the displacement between the items is greater than or equal to 45°. These images were generated with the baseline image parameters: $m = 4$, $n = 60$, $k = 2$. (*b–d*) Six example images show different choices of image parameters used in our experiment: item size (*b*), number of items (*c*) and image size (*d*), the size of an invisible central square in which items are randomly placed. All images shown here belong to *same* and *vertical* categories. When more than two items are used, the SD category label is determined by whether there are at least two identical items in the image. The SR category label is determined according to whether the average orientation of the displacements between all pairs of items is greater than or equal to 45°.

according to whether the *average* orientation of the displacements between all pairs of items is greater than or equal to 45°. Each image is generated by first drawing a joint class label for SD and SR from a uniform distribution over {*same, different*} × {*horizontal, vertical*}. The first item is sampled from a uniform distribution in $\{-1, 1\}^{m \times m}$. Then, if the sampled SD label is *same*, between 1 and $k - 1$ identical copies of the first item are created. If the sampled SD label is *different*, no identical copies are made. The rest of the k unique items are then consecutively sampled. These k items are then randomly placed in an $n \times n$ image while ensuring at least one background pixel spacing between items. Generating images by always drawing class labels for both problems ensures that the image distribution is identical between the two problem types.

We trained the same CNN repeatedly from scratch over multiple subsets of the data in order to see if learnability depends on the dataset's image parameters. CNNs were trained on 20 million images and training accuracy was sampled every 200 000 images. These samples were averaged across the length of a training run as well as over multiple trials for each condition, yielding a scalar measure of learnability called 'mean area under the learning curve' (mean ALC). The ALC is high when accuracy increases earlier and more rapidly throughout the course of training and/or when it converges to a higher final accuracy by the end of training.

First, we found a baseline architecture which could easily learn both same–different and spatial-relation PSVRT problems for one parameter configuration (item size $m = 4$, image size $n = 60$ and item number $k = 2$). Then, for a range of combinations of item size, image size and number of items, we trained an instance of this architecture from scratch. If a network learns the underlying rule of each visual relation, the resulting representations will be efficient at handling variations unrelated to the relation (e.g. a feature set to detect *any* pair of items arranged horizontally). As a result, the network should be equally good at learning the same problem in other image datasets with greater intra-category variability. In other words, the ALC will be consistently high over a range of image parameters. Alternatively, if the network's architecture does not allow for such representations and thus is only able to learn prototypes of examples within each category, the architecture will be progressively worse at learning the same visual relation instantiated with higher image variability. In this case, the ALC will gradually decrease as image variability increases.

The baseline CNN we used in this experiment had four convolutional layers. The first layer had eight filters with a 4×4 receptive field size. In the rest of the convolutional layers, filter size was fixed at 2×2 with the number of filters in each layer doubling from the immediately preceding layer. All convolutional layers had ReLU activations with strides of 1. Pooling layers were placed after every convolutional layer, with pooling kernels of size 3×3 and strides of 2. On top of retinotopic layers, all nine CNNs had three fully connected layers with 256 hidden units in each layer, followed by a two-dimensional classification layer. All network parameters were initialized using Xavier initialization [13] and were trained using the Adaptive Moment Estimation (Adam) optimizer [14] with a base learning rate of $\eta = 10^{-4}$. All experiments were run using TensorFlow [15]. To understand the effect of network size on learnability, we also used two control networks in this experiment: (i) a 'wide' control that had the same depth as the baseline but twice as many filters in the convolutional layers and four times as many hidden units in the fully connected layers and (ii) a 'deep' control which had twice as many convolutional layers as the baseline, by adding a convolutional layer of filter size 2×2 after each existing convolutional layer. Each extra convolutional layer had the same number of filters as the immediately preceding convolutional layer.

We varied each of three image parameters separately to examine its effect on learnability. This resulted in three sub-experiments (n was varied between 30 and 180, while m and k were fixed at 4 and 2, respectively; m was varied between 3 and 7, while n and k were fixed at 60 and 2, respectively; k was varied between 2 and 6, while n and m were fixed at 60 and 4, respectively). To use the same CNN architecture over a range of image sizes n, we fixed the actual input image size at 180×180 pixels by placing a smaller PSVRT image (if $n < 180$) at the centre of a blank background of size 180×180 pixels. The baseline CNN was trained from scratch in each condition with 20 million training images and a batch size of 50.

3.3. Results

In all conditions, we found a strong dichotomy in the observed learning curves. In cases where learning occurred, training accuracy abruptly jumped from chance level and gradually plateaued. We call this sudden, dramatic rise in accuracy the 'learning event'. When there was no learning event, accuracy remained at chance throughout a training session and the ALC

was 0.5. Strong bi-modality was observed even within a single experimental condition in which the learning event took place in only a subset of 10 randomly initialized trials. This led us to use two different quantities for describing a model's performance: (i) mean ALC obtained from *learned* trials (in which accuracy crossed 55%) and (ii) the number of trials in which the learning event never took place (*non-learned*). Note that these two quantities are independent, computed from two complementary subsets of 10 trials.

In SR, across all image parameters and in all trials, the learning event immediately occurred at the start of training and quickly approached 100% accuracy, producing consistently high and flat mean ALC curves (figure 5, blue dotted lines). In SD, however, we found that the overall ALC was significantly lower than SR (figure 5, red dotted lines).

In addition, we have also identified two main ways in which image variability affects learnability. First, among the trials in which the learning event did occur, the final accuracy achieved by the CNN at the end of training gradually decreased as the image size (n) of the number of items (k) increased. This caused the ALC to decrease from around 0.95 to 0.8. Second, increasing image size (n) also made the learning event decreasingly likely, with more than half of the trials failing to escape the chance level when image size was greater than 60 (figure 5, grey bars). We call this systematic degradation of performance accompanied by the increase in image variability the *straining effect*. In contrast, increasing item size produced no visible straining effect on the CNN. Similar to SR, learnability, in terms of both the frequency of the learning event as well as final accuracy, did not change significantly over the range of item sizes we considered.

The fact that straining is only observed in SD and not in SR and that it is only observed along some of the image parameters, n and k, suggests that straining is not simply a direct outcome of an increase in image variability. Using a CNN with more than twice the number of free parameters (figure 5, purple dotted lines) or with twice as many convolutional layers (figure 5, brown dotted lines) as a control did not qualitatively change the trend observed in the baseline model. Although increasing network size did result in improved learned accuracy in general, it also made learning less likely, yielding more non-learned trials than the baseline CNN.

We also rule out the possibility of the loss of spatial acuity from pooling or subsampling operations as a possible cause of straining for two reasons. First, our CNNs achieved the best overall accuracy when image size was smallest. If the loss of spatial acuity was the source of straining, increasing image size should have improved the network's performance instead of hurting it because items would have tended to be placed further apart from each other. Second, as we will show in experiment 3.2, an identical convolutional network where objects are forcibly separated into different channels does not exhibit any straining, suggesting that it is not the loss of spatial acuity per se that makes the SD problem difficult, but rather the fact that CNNs lack the ability to spatially separate representations of individual items in an image.

We hypothesize that these straining effects reflect the way the positioning of each item contributes to image variability. A little arithmetic shows that image variability is an exponential function of image size as the base and number of items as the exponent. Thus, increasing image size while fixing the number of items at two results in a quadratic-rate increase in image variability, while increasing the number of items leads to an exponential-rate increase in image variability. Image variability is also an exponential function of item size as the exponent and 2 (for using binary pixels) as the base.

The comparatively weak effects of item size and item number shed light on the computational strategy used by CNNs to solve

SD. Our working hypothesis is that CNNs learn 'subtraction templates', filters with one positive region and one negative region (like a Haar or Gabor wavelet), in order to detect the similarity between two image regions. A different subtraction template is required for each relative arrangement of items, since each item must lie in one of the template's two regions. When identical items lie in these opposing regions, they are effectively subtracted by the synaptic weights. This difference is then used to choose the appropriate same–different label. Note that this strategy does not require memorizing specific items. Hence, increasing item size (and therefore total number of possible items) should not make the task appreciably harder. Further, a single subtraction template can be used even in scenes with more than two items, since images are classified as 'same' when they have *at least* two identical items. So, any straining effect from item number should be negligible as well. Instead, the principal straining effect with this strategy should arise from image size, which increases the possible number of arrangements of items.

Taken together, these results suggest that, when CNNs learn a PSVRT condition, they are simply building a feature set tailored to the relative positional arrangements of items in a particular dataset, instead of learning the abstract 'rule' per se. If a network is able to learn features that capture the visual relation at hand, then these features should, by definition, be minimally sensitive to the image variations that are irrelevant to the relation. This seems to be the case only in SR. In SD, increasing image variability lowered the ALC for the CNNs. This suggests that the features learned by CNNs are not invariant rule detectors, but rather merely a collection of templates covering a particular distribution in the image space.

4. Experiment 3: Is object individuation needed to solve visual relations?

Our main hypothesis is that CNNs struggle to learn visual relations in part because they are feedforward architectures which lack a mechanism for grouping features into individuated objects. Recently, however, Santoro *et al.* [11] proposed the relational network (RN), a feedforward architecture aimed at learning visual relations without such an individuation mechanism. RNs are fully connected feedforward networks which operate on pairs of so-called 'objects' (figure 6; for concision, we will refer to a neural network consisting of a CNN feeding into an RN as just an RN). These objects are simply feature columns from all retinotopic locations in a deep layer of a CNN, similar to the feature columns found in higher areas of the visual cortex [17]. These feature vectors will sometimes represent parts of the background, incomplete items or even multiple items because the network does not explicitly represent individual objects. This makes the 'objects' used by an RN rather different from those discussed in the psychophysical literature, where perceptual objects are speculated to obey gestalt rules like boundedness and continuity [18]. Santoro *et al.* [11] emphasize that their model performed well even though it employs this highly unstructured notion of object: 'A central contribution of this work is to demonstrate the flexibility with which relatively unstructured inputs, such as CNN or LSTM [long short-term memory] embeddings, can be considered as a set of objects for an RN.'

In particular, the RN was able to outperform a baseline CNN on the 'sort-of-CLEVR' challenge, a visual question-answering task using images with simple geometric items (see figure 7a for examples of sort-of-CLEVR items).
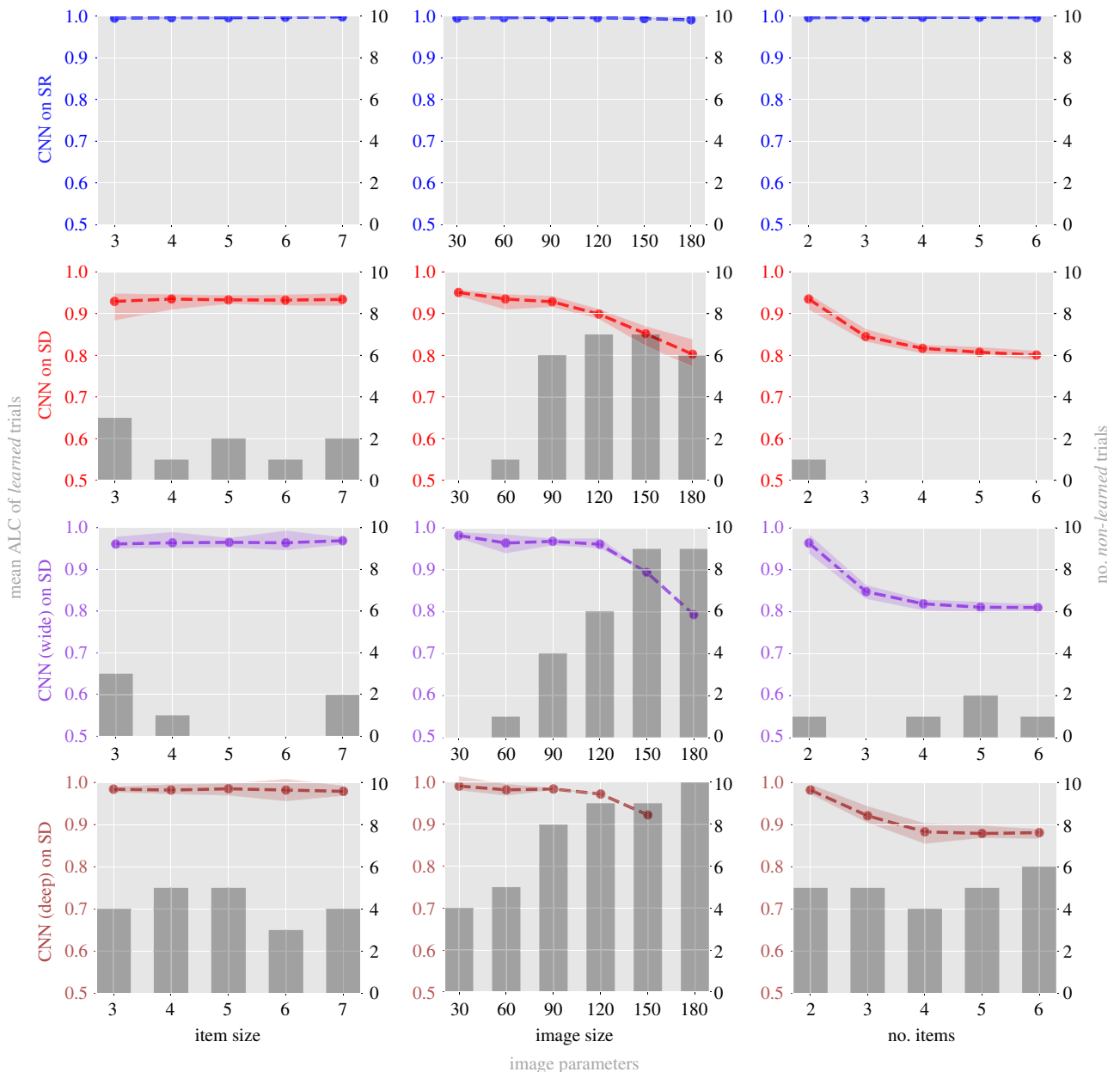
Figure 5. Mean area under the learning curve (ALC) over PSVRT image parameters. ALC is the normalized area under a training accuracy curve over the course of training on 20 million images. Coloured dots are the mean ALCs of learned trials (trials in which validation accuracy exceeded 55%) out of 10 randomly initialized trials. Shaded regions around the coloured dots indicate the intervals between the maximum and the minimum ALC among learned trials. Grey bars denote the number of non-learned trials, out of 10 trials, in which validation accuracy never exceeded 55%. Three model – task combinations (CNN on SR (blue), CNN on SD (red), wide CNN control on SD (purple) and deep CNN control on SD (brown)) are plotted, and each combination is explored over three image variability parameters: item size, image size and number of items.

In sort-of-CLEVR, scenes contain up to six items, each of which has one of two shapes and six colours. The RN was trained to answer both relational questions (e.g. *What is the shape of the object that is furthest from the grey object?*) and non-relational questions (e.g. *Is the red object on the top or bottom of the scene?*).

However, the sort-of-CLEVR task suffers from three important shortcomings. First, the number of possible items is exceedingly small (6 colours × 2 shapes = 12 items). Combined with the fact that the authors used rather small (75 × 75) images, this means the total number of sort-of-CLEVR stimuli was rather low, at least compared with PSVRT stimuli. The small number of samples in sort-of-CLEVR might have encouraged the RN to use rote memorization instead of actually learning relational concepts. Second, while the authors trained the

RN to compare the attributes of scene items (e.g. *How many objects have the same shape as the green object?*), they did not examine if the model could learn the concept of sameness, per se (e.g. *Are any two items the same in this scene?*). Detecting sameness is a particularly hard task because it requires matching all attributes between all pairs of items. Third, sort-of-CLEVR stimuli are not parameterized as they are in PSVRT; one cannot systematically vary image features while keeping the abstract rule fixed. Thus, it is difficult to say whether the success of RNs arises from their ability to flexibly learn relations among arbitrary objects (as is hypothesized for humans [19]) or rather their ability to fit particular image features.

Crucially, without a parameterized dataset, it is difficult to evaluate the authors' claim regarding the efficacy of 'relatively unstructured' objects in visual reasoning problems.
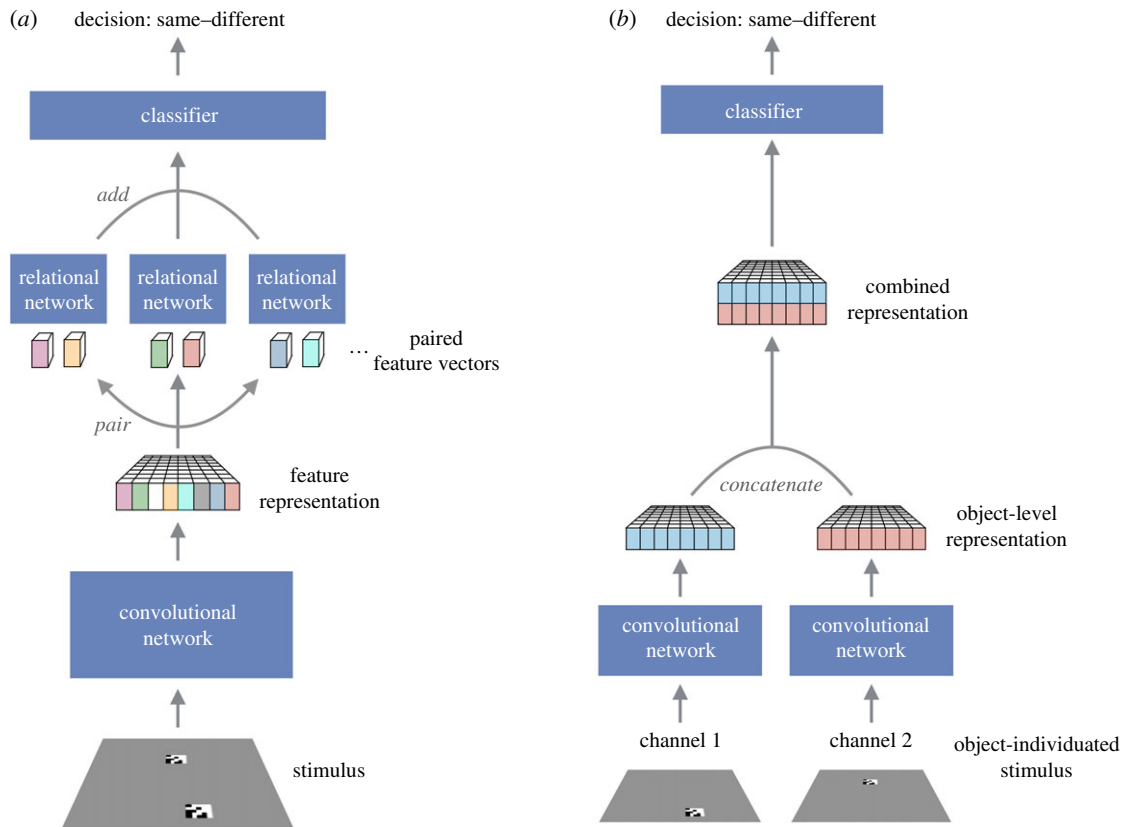
Figure 6. A comparison between a relational network and the proposed Siamese architecture. (*a*) A relational network ((*a*), top half) is a fully connected, feedforward neural network which accepts pairs of CNN feature vectors as input. First, the image is passed through a CNN to extract features. Every pair of feature activations ('objects') at every retinotopic location in the final CNN layer is passed through the RN. The outputs of the RN on every pair of activations is then summed and passed through a final feedforward network, producing the decision. Depending on the spatial resolution of the final CNN layer and the receptive field of each unit, the object representations of an RN may correspond to a single scene item, multiple items, partial items or even the background. (*b*) In contrast, objects in our Siamese network are forced to contain a single item. First, we split stimuli into several images, each containing a single item. Then, each of the images is passed through a separate CNN (here, channel 1 and channel 2), producing a representation of a single object. These objects are then combined by concatenation into a single representation and passed through a classifier. The network simulates the effects of the attentional and perceptual grouping processes suspected to underlie biological visual reasoning (see Discussion).

Since the objects used by RNs are simply feature columns, they have a fixed receptive field. Thus, the success of RNs on sort-of-CLEVR might be due to felicitously sized and arranged items instead of actual relational learning. For, if image features are allowed to parametrically vary, such spatially rigid representations might fail to correctly encode individual objects whenever, for instance, multiple, small and tightly arranged items fall within the same receptive field or when a large, irregularly shaped item spans multiple receptive fields.

Our goal in experiment 3 was to re-evaluate relational networks on sort-of-CLEVR when these handicaps are removed. To that end, we performed three sub-experiments. First, we trained RNs on a bona fide same–different task using versions of sort-of-CLEVR missing certain colour–shape combinations in order to see if the model would over-fit to training item attributes (see [20] for a similar demonstration in a different visual reasoning problem). Such over-fitting would indicate that the RN merely memorizes particular item combinations instead of learning abstract rules. Second, we tested an RN on PSVRT in order to evaluate the ease with which the model can fit data when scene items systematically vary in appearance and arrangement. As in experiment 2, we measured the mean ALC in order to see if the RN's object representations alleviated the straining found in CNNs.

Finally, we compared the performance of the RN on PSVRT with that of an idealized model using ground-truth object individuation. Our new model is a 'Siamese' network [21] which processes each scene item in a separate (CNN) channel and then passes the processed items to a single classifier network. This model simulates the effects of attentional selection and perceptual grouping by segregating the representations of each item. Unlike an RN, whose object representations may in fact contain no items, multiple items or incomplete items, object representations in the Siamese network contain exactly one item.

4.1. Methods
4.1.1. Sub-experiment 3.1: Relational transfer to novel attribute combinations
Here, we sought to measure the ability of an RN to transfer the concept of sameness from a training set to a novel set of objects, a classic and very-well-studied paradigm in animal psychology (see [22] for a review) and thus an important benchmark for models of visual reasoning. We used software for relational networks publicly available at https://github.com/gitlimlab/Relation-Network-Tensorflow. Like the original architecture used by Santoro *et al.* [11], our RN had four convolutional layers with ReLU non-linearities and batch normalization. We used 24 features for each convolutional layer, fewer than
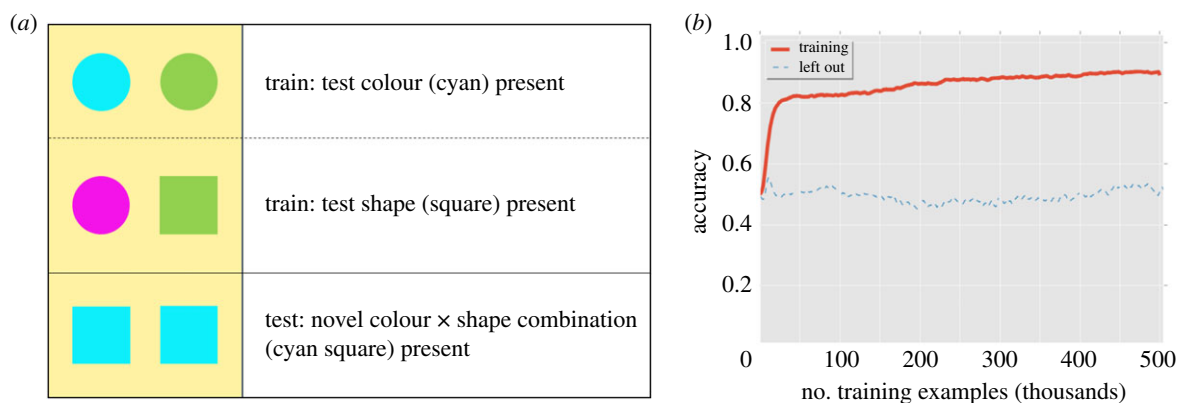
Figure 7. (a) Sample items used during training and testing in experiment 3. We trained relational networks (RNs) on 12 two-item same−different datasets each missing one colour−shape combination from sort-of-CLEVR (2 shapes × 6 colours). Then, we tested the model on the left-out combination. The top and middle rows of panel (a) show two possible pairs of items when the left-out combination is 'cyan square'. Row 1 shows a cyan circle and row 2 shows a green square. However, only in the test set is the model queried about images involving a cyan square (e.g. the 'same' image in row 3). Note that, during training, the model observes each left-out attribute, just not in the left-out combination. (b) Averaged accuracy curves of an RN while being trained on the sort-of-CLEVR datasets missing one colour−shape combination. The red curve shows the training accuracy. The blue dashed line shows the accuracy on validation data with the left-out items.

those used by [11], but sufficient for good training accuracy. These convolutional layers were followed by two four-layer multi-layer perceptrons (MLPs), both with ReLU non-linearities. These MLPs had 256 features each, again fewer than those in [11], but sufficient for fitting the data. The final classification layer had a softmax nonlinearity and the whole network was optimized with a cross-entropy loss using an Adam optimizer with learning rate $\eta = 10^{-4}$ and mini-batches of size 64. The original authors did not report receptive field sizes or strides. Our RN used receptive field sizes of 5×5 throughout the convolutional layers and had strides of 3 in the first two convolutional layers and strides of 2 in the next two. There was no pooling. We confirmed that this model was able to reproduce the results from [11] on the sort-of-CLEVR task.

We constructed 12 different versions of the sort-of-CLEVR dataset, each one missing one of the 12 possible colour × shape attribute combinations (see figure 7a). Images in each dataset only depicted two items, randomly placed on a 128×128 background. Half of the time, these items were the same (same colour and same shape). For each dataset, we trained the RN architecture to detect the possible sameness of the two scene items while measuring validation accuracy on the left-out images. We then averaged training accuracy and validation accuracy across all of the left-out conditions.

4.1.2. Sub-experiment 3.2: Relational networks on PSVRT

For this experiment, we trained an RN on our experiment 2 with PSVRT stimuli, and observed whether the straining effect found in CNNs was alleviated in RNs. For this sub-experiment, we used the exact architecture from sub-experiment 3.1, but increased the number of units to the original values from [11] in order to give the RN the best possible chance of learning the very difficult PSVRT task. The convolutional layers had 32, 64, 128 and 256 features, the first MLP had 2000 units in each layer, and the final MLP had 2000, 1000, 5000 and 100 units in its four layers. We focused only on same−different learning and only varied the image size from 30 to 180 pixels since this produced the strongest straining effect in CNNs. Item size was fixed at 4 and the number of items was fixed at 2. We trained on 20 million images, using 10 randomly initialized trials. As in experiment 2, we measured the mean ALC

as well as the number of non-learned trials. Before training on the whole spectrum of image sizes, we ensured that the RN was capable of fitting the data when item size was 4 and image size was 60.

4.1.3. Sub-experiment 3.3: The need for perceptual grouping and object individuation

Here, we introduce a Siamese network which processes scene items individually in separate CNN 'channels' (figure 6b). First, we manually split each PSVRT stimulus into several images, each of which contained a single item. These images were then individually processed by two copies of the same network (mimicking, in a sense, the process of sequentially attending to individuated objects). For example, if one stimulus contained two objects in the original PSVRT, our new stimulus would be presented to the Siamese network as two separate images. The scene items retained their original location in each image so that item position varied just as widely as in the original PSVRT. These images were then individually processed by each CNN channel, using the same architecture as in experiment 2. This resulted in two object-separated feature maps in the topmost retinotopic layer (figure 6b). These feature maps were then concatenated before being passed to the fully connected classifier layers.

This Siamese configuration is essentially an idealized version of the kinds of object representations resulting from psychological processes such as perceptual grouping and attentional selection. Because convolutional layers in this configuration are now constrained to process only one object at a time, regardless of the total number of objects presented in an image, the network can completely disregard the positional information of individual objects and only preserve information about their identities under comparison.

4.2. Results
4.2.1. Sub-experiment 3.1: Relational transfer to novel attribute combinations

In the sort-of-CLEVR transfer task, we found that the RN does not generalize on average to left-out colour−shape

attribute combinations (figure 7). Since there are only 11 colour–shape combinations in any given set-up, the model did not need to learn to generalize across many items. As a result, the RN learned orders of magnitude faster than the CNNs in experiment 2; for example, average training accuracy (solid red) exceeded 80% within 50 000 examples. However, while the average training accuracy curve rose rapidly to around 90%, the average validation accuracy remained at chance. In other words, there was no transfer of same–different ability to the left-out condition, even though the attributes from that condition (e.g. cyan square) were represented in the training set, just not in that combination (e.g. cyan circle and green square; figure 7a).

4.2.2. Sub-experiment 3.2: Relational networks on PSVRT

We found that the RN exhibits a qualitatively similar straining effect to increasing image size (figure 8, pale blue dotted lines). Similar to CNNs in experiment 2, the mean ALC of learned trials gradually decreased as image size increased, together with the observed likelihood of learning out of 10 restarts. Since the top retinotopic feature vectors that are treated as 'object representations' in RN have rather large, fixed and highly overlapping receptive fields, the RN is strained just as easily as regular CNNs. In order to accommodate this fixed architecture, the RN must learn a dictionary of features that captures all arrangements of items for a given image size condition. This is an increasingly difficult feat as the image size grows, straining the model heavily until it simply cannot learn at all in the final condition (image size 180×180).

4.2.3. Sub-experiment 3.3: The need for perceptual grouping and object individuation

The mean ALC curves for the Siamese network on PSVRT were strikingly different from those of the CNN in experiment 2 (figure 8, first and second rows). Barely any straining effect was observed on the SD task, and the model learned within 5 million examples across all image size parameters in either the SD or SR tasks. In SD, since objects are individuated by fiat, the network need not learn all possible spatial arrangements of items. The network must simply learn to compare whichever two items reach the classifier layers through the two CNN channels. This greatly simplifies the SD problem, alleviating straining. In both SD and SR, the Siamese network can learn to flexibly represent the task-relevant properties of each object such that learnability is not at all influenced by image variability. In other words, a feedforward network, once endowed with object individuation, can easily construct invariant feature representations with which arbitrary objects can be related.

This result implies that object individuation makes visual relation detection a rather trivial problem for feedforward networks. In informal experiments (data not shown) we found that even very shallow Siamese networks (e.g. with one convolutional layer) could still learn SD much faster than baseline CNNs. Naturally, we do not intend our Siamese network as a bona fide solution to visual reasoning, but rather as a proof of the efficacy of object individuation in visual reasoning problems. A genuine visual reasoning model would be able to dynamically select and group features in the scene (see Discussion section).

5. Discussion

Recent progress in computational vision has been significant [23]. Modern deep learning architectures can discriminate between 1000 object categories [3] and identify faces among millions of distractors [24] at a level approaching—and possibly even surpassing—that of human observers. While these neural networks do not aim to mimic the organization of the visual cortex in detail, they are at least partly inspired by biology. Modern deep learning architectures are indeed closely related to earlier hierarchical models of the visual cortex albeit with much better categorization accuracy (see [25,26] for reviews). Further, CNNs have been shown to account well for monkey inferotemporal data [27] and human lateral occipital data [28,29]. In addition, deep networks have been shown to be consistent with a number of human behaviours including rapid visual categorization [30,31], image memorability [32], typicality [33] as well as similarity [34] and shape sensitivity [35] judgements.

Concurrently, a growing body of literature has been highlighting key dissimilarities between current deep network models and various aspects of visual cognition. One prominent example is adversarial perturbation [36], a type of structured image distortion that asymmetrically affects CNNs and humans. Although barely perceptible to a human observer, adversarial perturbation renders an image unrecognizable to a CNN, even though the same CNN can correctly recognize the unperturbed image with high confidence. Another example is the poor generalization of CNNs in conditions that pose no difficulty to human observers, such as learning novel object categories with minimal supervision or when the parts of a familiar object are shown in unfamiliar but realistic configurations [37–39]. Direct evidence for qualitatively different feature representations used by humans and CNNs was shown in [40,41].

The present study adds to this body of literature by demonstrating feedforward neural networks' fundamental inability to efficiently and robustly learn visual relations. Our results indicate that visual-relation problems can quickly exceed the representational capacity of feedforward networks. While learning feature templates for single objects appears tractable for modern deep networks, learning feature templates for *arrangements* of objects becomes rapidly intractable because of the combinatorial explosion in the requisite number of templates. That notions of 'sameness' and stimuli with a combinatorial structure are difficult to represent with feedforward networks has long been acknowledged by cognitive scientists [42,43].

Compared with the feedforward networks in this study, biological visual systems excel at detecting relations. Fleuret *et al.* [5] found that human observers are capable of learning rather complicated visual rules and generalizing them to new instances from just a few training examples. Participants could learn the rule underlying the hardest SVRT problem for CNNs in our experiment 1, problem 20, from an average of about six examples. Problem 20 is rather complicated as it involves two shapes such that '*one shape can be obtained from the other by reflection around the perpendicular bisector of the line joining their centers*' ([5], fig. S26, suppl. p. 27). In contrast, the best performing CNN model for this problem could not get significantly above chance from 1 million training examples.
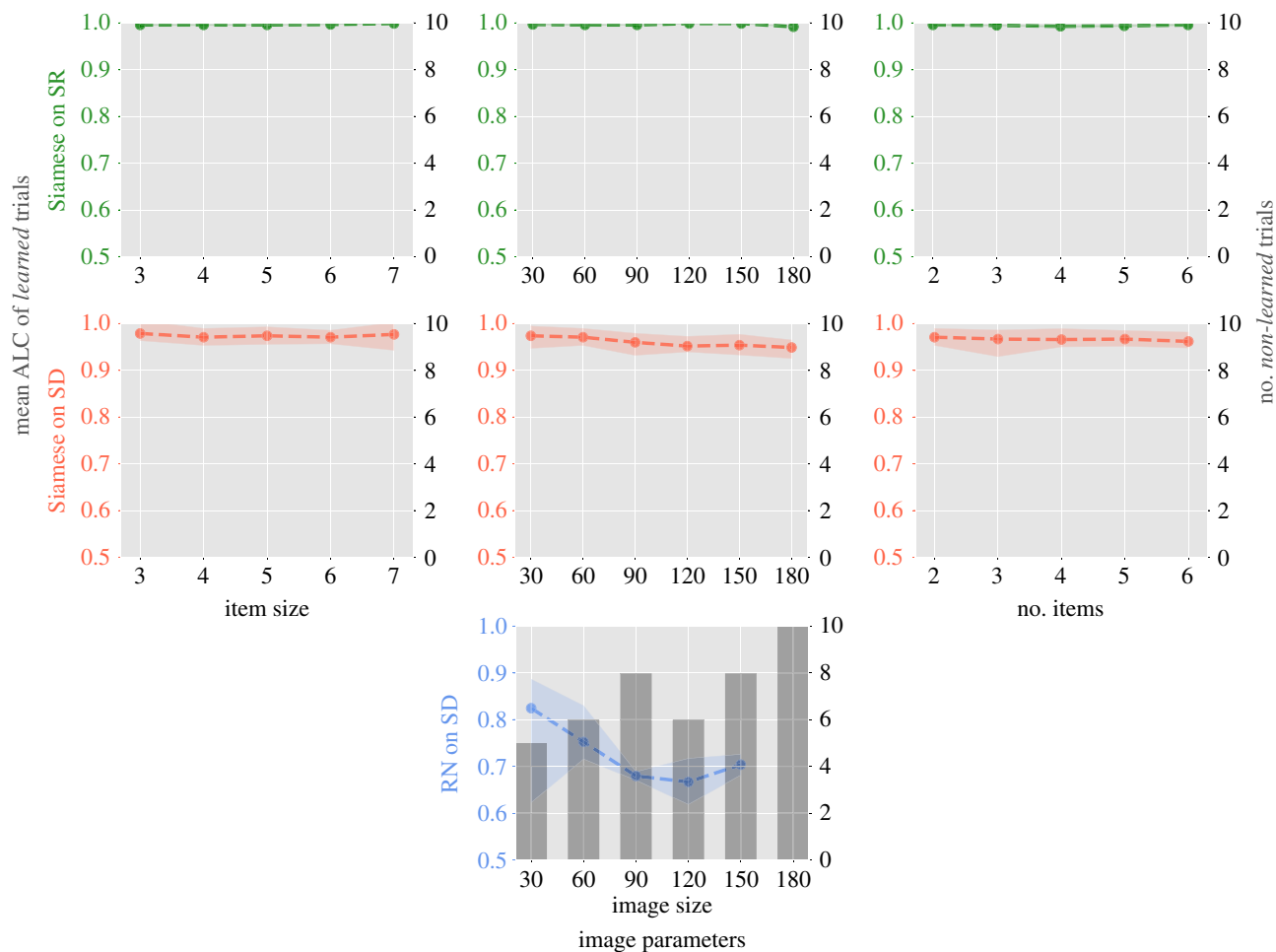
Figure 8. Mean ALC of the Siamese network on SD and SR tasks and the RN on SD over image sizes. Unlike for CNNs, mean ALC curves of the Siamese network exhibit no significant straining. The network learns equally well on all datasets with different image variability parameters. The significant difference between SD and SR conditions observed with CNNs is no longer present in the Siamese network. In contrast, the RN exhibits a strong straining effect that is qualitatively similar to CNNs, with the average ALC as well as the probability of learning decreasing as image size increases.

This failure of modern computer vision algorithms is all the more striking given the widespread ability to recognize visual relations across the animal kingdom. Previous studies showed that non-human primates [44,45], birds [2,46], rodents [47] and even insects [48] can be trained to recognize abstract relations between training objects and then transfer this knowledge to novel objects. Contrast the behaviour of the ducklings in [2] with the RN of experiment 3, which demonstrated no ability to transfer the concept of same–different to novel objects (figure 7) even after hundreds of thousands of training examples.

There is substantial evidence that visual-relation detection in primates depends on re-entrant/feedback signals beyond feedforward, pre-attentive processes. It is relatively well accepted that, despite the widespread presence of feedback connections in our visual cortex, certain visual recognition tasks, including the detection of natural object categories, are possible in the near absence of cortical feedback—based primarily on a single feedforward sweep of activity through our visual cortex [49]. However, psychophysical evidence suggests that this feedforward sweep is too spatially coarse to localize objects even when they can be recognized [50]. The implication is that object localization in clutter requires attention [51]. It is difficult to imagine how one could recognize a relation between two objects without spatial information. Indeed, converging evidence [19,52–56] suggests that the processing of spatial relations between pairs of objects in a cluttered scene requires attention, even when individual objects can be detected pre-attentively.

Another brain mechanism implicated in our ability to process visual relations is working memory [57–60]. In particular, imaging studies [57,58] have highlighted the role of working memory in prefrontal and pre-motor cortices when participants solve Raven's progressive matrices which require both spatial and same–different reasoning.

What is the computational role of attention working memory in the detection of visual relations? One assumption [19] is that these two mechanisms allow flexible representations of relations to be constructed *dynamically* at run-time via a sequence of attention shifts rather than *statically* by storing visual-relation templates in synaptic weights (as done in feedforward neural networks). Such representations built 'on-the-fly' circumvent the combinatorial explosion associated with the storage of templates for all possible relations, helping to prevent the capacity overload that plagues feedforward neural networks.

Humans can easily recognize when two objects are the same up to some transformation [1] or when objects exist in a given spatial relation [5,19]. More generally, humans can effortlessly construct an unbounded set of structured descriptions about their visual world [61]. Mechanisms in the visual system such as perceptual grouping, attention and working memory exemplify how the brain learns and

handles combinatorial structures in the visual environment with small amount of experience [62]. However, exactly how attentional and mnemonic mechanisms interact with hierarchical feature representations in the visual cortex is not well understood. Given the vast superiority of humans over modern computers in their ability to detect visual relations, we see the exploration of these cortical mechanisms as a crucial step in our computational understanding of visual reasoning.

Endnote

[1]A shorter version [12] of this paper is to appear in the proceedings of the 40th Annual Conference of the Cognitive Science Society.

References

1. Shepard RN, Metzler J. 1971 Mental rotation of three-dimensional objects. *Science* **171**, 701–703. (doi:10.1126/science.171.3972.701)

2. Martinho III A, Kacelnik A. 2016 Ducklings imprint on the relational concept of 'same or different'. *Science* **353**, 286–288. (doi:10.1126/science.aaf4247)

3. He K, Zhang X, Ren S, Sun J. 2015 Delving deep into rectifiers: surpassing human-level performance on imagenet classification. (http://arxiv.org/abs/1502.01852).

4. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L 2009 ImageNet: A large-scale hierarchical image database. In *Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 20–25 June 2009*, pp. 248–255. New York, NY: IEEE.

5. Fleuret F, Li T, Dubout C, Wampler EK, Yantis S, Geman D. 2011 Comparing machines and humans on a visual categorization test. *Proc. Natl Acad. Sci. USA* **108**, 17 621–17 625. (doi:10.1073/pnas.1109168108)

6. LeCun Y, Bottou L, Bengio Y, Haffner P. 1998 Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2323. (doi:10.1109/5.726791)

7. Krizhevsky A, Sutskever I, Hinton GE. 2012 ImageNet classification with deep convolutional neural networks. In *Proc. NIPS'12: 25th Int. Conf. on Neural Information Processing Systems, Lake Tahoe, NV, 3–6 December 2012*, pp. 1097–1105. Red Hook, NY: Curran Associates.

8. Ellis K, Solar-lezama A, Tenenbaum JB 2015 Unsupervised learning by program synthesis. In *Advances in neural information processing systems*, vol. 28, pp. 973–981. New York, NY: Curran Associates, Inc.

9. Stabinger S, Rodríguez-Sánchez A, Piater J. 2016 25 years of CNNs: can we compare to human abstraction capabilities? *ICANN* **9887**, 380–387.

10. Gülçehre Ç, Bengio Y. 2013 Knowledge matters: importance of prior information for optimization. (http://arxiv.org/abs/1301.4083).

11. Santoro A, Raposo D, Barrett DGT, Malinowski M, Pascanu R, Battaglia P, Lillicrap T. 2017 A simple neural network module for relational reasoning. (http://arxiv.org/abs/1706.01427).

12. Ricci M, Kim J, Serre T. In press. Same-different problems strain convolutional neural networks. In *Proc. of the 40th Annu. Meeting of the Cognitive Science Society*. Madison, WI: Cognitive Science Society.

13. Glorot X, Bengio Y. 2010 Understanding the difficulty of training deep feedforward neural networks. In *Proc. of the 13th Int. Conf. on Artificial Intelligence and Statistics* (eds YW Teh, M Titterington). Proceedings of Machine Learning Research, vol. 9, pp. 249–256. Fort Lauderdale, FL: PMLR.

14. Kingma DP, Ba J. 2015 Adam: A method for stochastic optimization. In *Proc. of the 3rd Int. Conf. on Learning Representations, San Diego, CA, 7–9 May 2015*.

15. Abadi M et al. 2016 TensorFlow: a system for large-scale machine learning. In *Proc. of the 12th USENIX Conf. on Operating Systems Design and Implementation, Savannah, GA, 2–4 November 2016*, pp. 265–283. Berkeley, CA: USENIX Association.

16. Stabinger S, Rodriguez-Sanchez A. 2017 Evaluation of deep learning on an abstract image classification dataset. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV) Workshops, Venice, Italy, 22–29 October 2017*. New York, NY: IEEE

17. Tanaka K. 2003 Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities. *Cereb. Cortex* **13**, 90–99. (doi:10.1093/cercor/13.1.90)

18. Spelke ES, Katz G, Purcell SE, Ehrlich SM, Breinlinger K. 1994 Early knowledge of object motion: continuity and inertia. *Cognition* **51**, 131–176. (doi:10.1016/0010-0277(94)90013-2)

19. Franconeri SL, Scimeca JM, Roth JC, Helseth SA, Kahn LE. 2012 Flexible visual processing of spatial relationships. *Cognition* **122**, 210–227. (doi:10.1016/j.cognition.2011.11.002)

20. Johnson J, Hariharan B, van der Maaten L, Fei-Fei L, Zitnick CL, Girshick R. 2017 CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 21–26 July 2017*, pp. 1988–1997. Washington, DC: IEEE.

21. Bromley J, Guyon I, LeCun Y, Säckinger E, Shah R. 1993 Signature verification using a 'siamese' time delay neural network. In *Advances in neural information processing systems*, vol. 6, pp. 737–744. Cambridge, MA: MIT Press.

22. Wright AA, Kelly DM. 2017 Comparative approaches to same/different abstract concept-learning. *Learn. Behav.* **45**, 323–324. (doi:10.3758/s13420-017-0271-4)

23. LeCun Y, Bengio Y, Hinton G. 2015 Deep learning. *Nature* **521**, 436–444. (doi:10.1038/nature14539)

24. Kemelmacher-Shlizerman I, Seitz SM, Miller D, Brossard E. 2016 The megaface benchmark: 1 million faces for recognition at scale. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, 27–30 June 2016*, pp. 4873–4882. New York, NY: IEEE.

25. Serre T. 2015 Hierarchical models of the visual system. In *Encyclopedia of computational neuroscience* (eds D Jaeger, R Jung), pp. 1309–1318. New York, NY: Springer.

26. Kriegeskorte N. 2015 Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* **1**, 417–446. (doi:10.1146/annurev-vision-082114-035447)

27. Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. 2014 Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 8619–8624. (doi:10.1073/pnas.1403112111)

28. Khaligh-Razavi S-M, Kriegeskorte N. 2014 Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915. (doi:10.1371/journal.pcbi.1003915)

29. Guclu U, van Gerven MAJ. 2015 Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10 005 – 10 014. (doi:10.1523/JNEUROSCI.5023-14.2015)

30. Kheradpisheh SR, Ghodrati M, Ganjtabesh M, Masquelier T. 2016 Deep networks can resemble human feed-forward vision in invariant object recognition. *Sci. Rep.* **6**, 32672. (doi:10.1038/srep32672)

31. Eberhardt S, Cader JG, Serre T. 2016 How deep is the feature analysis underlying rapid visual categorization? In *Advances in neural information processing systems*, vol. 29, pp. 1100 – 1108. New York, NY: Curran Associates, Inc.

32. Dubey R, Peterson J, Khosla A, Yang M-H, Ghanem B. 2015 What makes an object memorable? In *Proc. of the IEEE Int. Conf. on Computer Vision, Santiago, Chile, 7 – 13 December 2015*, pp. 1089 – 1097. Washington, DC: IEEE.

33. Lake BM, Zaremba W, Fergus R, Gureckis TM. 2015 Deep neural networks predict category typicality ratings for images. In *Proc. of the 37th Annual Conf. of the Cognitive Science Society, Pasadena, CA, 22 – 25 July 2015*. Austin, Texas: Cognitive Science Society.

34. Peterson J, Abbott J, Griffiths T. 2016 Adapting deep network features to capture psychological representations. *Proc. of the 26th Int. Joint Conference on Artificial Intelligence (IJCAI), Melbourne, Australia, 19 – 25 August 2017*, pp. 4934 – 4938. International Joint Conferences on Artificial Intelligence.

35. Kubilius J, Bracci S, Op de Beeck HP. 2016 Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput. Biol.* **12**, e1004896 – 00. (doi:10.1371/journal.pcbi.1004896)

36. Goodfellow IJ, Shlens J, Szegedy C. 2015 Explaining and harnessing adversarial examples. In *Proc. Int. Conf. on Learning Representations, San Diego, CA, 7 – 9 May 2015*.

37. Lake B, Salakhutdinov R, Tenenbaum J. 2015 Human-level concept learning through probabilistic program induction. *Science* **350**, 1332 – 1338. (doi:10.1126/science.aab3050)

38. Saleh B, Elgammal A, Feldman J 2016 The role of typicality in object classification: improving the generalization capacity of convolutional neural networks. (https://arxiv.org/abs/1602.02865v1)

39. Erdogan G, Jacobs RA. 2017 Visual shape perception as Bayesian inference of 3D object-centered shape representations. *Psychol. Rev.* **124**, 740 – 761. (doi:10.1037/rev0000086)

40. Ullman S, Assif L, Fetaya E, Harari D. 2016 Atoms of recognition in human and computer vision. *Proc. Natl Acad. Sci. USA* **113**, 2744 – 2749. (doi:10.1073/pnas.1513198113)

41. Linsley D, Eberhardt S, Sharma T, Gupta P, Serre T. 2017 What are the visual features underlying human versus machine vision? In *Proc. IEEE Int. Conf. on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, 22 – 29 October 2017*. Washington, DC: IEEE Computer Society.

42. Fodor JA, Pylyshyn ZW. 1988 Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**, 3 – 71. (doi:10.1016/0010-0277(88)90031-5)

43. Marcus G. 2001 *The algebraic mind: integrating connectionism and cognitive science*. Cambridge, MA: MIT Press.

44. Donderi D, Zelnicker D. 1969 Parallel processing in visual same-different. *Percept. Psychophys.* **5**, 197 – 200. (doi:10.3758/BF03210537)

45. Katz JS, Wirght AA. 2006 Same/different abstract-concept learning by pigeons. *J. Exp. Psychol. Anim. Behav. Process.* **32**, 80 – 86. (doi:10.1037/0097-7403.32.1.80)

46. Daniel TA, Wright AA, Katz JS. 2015 Abstract-concept learning of difference in pigeons. *Anim. Cogn.* **18**, 831 – 837. (doi:10.1007/s10071-015-0849-1)

47. Wasserman EA, Castro L, Freeman JH. 2012 Same-different categorization in rats. *Learn. Mem.* **19**, 142 – 145. (doi:10.1101/lm.025437.111)

48. Giurfa M, Zhang S, Jenett A, Menzel R, Srinivasan MV. 2001 The concepts of 'sameness' and 'difference' in an insect. *Nature* **410**, 930 – 933. (doi:10.1038/35073582)

49. Serre T. 2016 Models of visual categorization. *Wiley Interdiscip. Rev. Cogn. Sci.* **7**, 197 – 213. (doi:10.1002/wcs.1385)

50. Evans KK, Treisman A. 2005 Perception of objects in natural scenes: is it really attention free?. *J. Exp. Psychol. Hum. Percept. Perform.* **31**, 1476 – 1492. (doi:10.1037/0096-1523.31.6.1476)

51. Zhang Y, Meyers EM, Bichot NP, Serre T, Poggio T, Desimone R. 2011 Object decoding with attention in inferior temporal cortex. *Proc. Natl Acad. Sci. USA* **108**, 8850 – 8855. (doi:10.1073/pnas.1100999108)

52. Logan GD. 1994 Spatial attention and the apprehension of spatial relations. *J. Exp. Psychol. Hum. Percept. Perform.* **20**, 1015 – 1036. (doi:10.1037/0096-1523.20.5.1015)

53. Moore CM, Elsinger CL, Lleras A. 1994 Visual attention and the apprehension of spatial relations: the case of depth. *J. Exp. Psychol. Hum. Percept. Perform.* **20**, 1015 – 1036. (doi:10.1037/0096-1523.20.5.1015)

54. Rosielle LJ, Crabb BT, Cooper EE. 2002 Attentional coding of categorical relations in scene perception: evidence from the flicker paradigm. *Psychon. Bull. Rev.* **9**, 319 – 326. (doi:10.3758/BF03196288)

55. Holcombe AO, Linares D, Vaziri-Pashkam M. 2011 Perceiving spatial relations via attentional tracking and shifting. *Curr. Biol.* **21**, 1135 – 1139. (doi:10.1016/j.cub.2011.05.031)

56. van der Ham IJM, Duijndam MJA, Raemaekers M, van Wezel RJA, Oleksiak A, Postma A. 2012 Retinotopic mapping of categorical and coordinate spatial relation processing in early visual cortex. *PLoS ONE* **7**, e38644. (doi:10.1371/journal.pone.0038644)

57. Kroger JK, Sabb FW, Fales CL, Bookheimer SY, Cohen MS, Holyoak KJ. 2002 Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: a parametric study of relational complexity. *Cereb. Cortex* **12**, 477 – 485. (doi:10.1093/cercor/12.5.477)

58. Golde M, von Cramon DY, Schubotz RI. 2010 Differential role of anterior prefrontal and premotor cortex in the processing of relational information. *Neuroimage* **49**, 2890 – 2900. (doi:10.1016/j.neuroimage.2009.09.009)

59. Clevenger PE, Hummel JE. 2014 Working memory for relations among objects. *Attention Perception Psychophys.* **76**, 1933 – 1953. (doi:10.3758/s13414-013-0601-3)

60. Brady TF, Alvarez GA. 2015 Contextual effects in visual working memory reveal hierarchically structured memory representations. *J. Vis.* **15**, 1 – 69. (doi:10.1167/15.15.6)

61. Geman D, Geman S, Hallonquist N, Younes L. 2015 Visual Turing test for computer vision systems. *Proc. Natl Acad. Sci. USA* **112**, 3618 – 3623. (doi:10.1073/pnas.1422953112)

62. Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND. 2011 How to grow a mind: statistics, structure, and abstraction. *Science* **331**, 1279 – 1285. (doi:10.1126/science.1192788)