Uncorrected draft. Manuscript to appear in Interface Focus theme issue “Understanding images in biological and computer vision”

Learning same-different relations strains feedforward neural networks

Junkyung Kim† Matthew Ricci† Thomas Serre

†equal contributions

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

Abstract

Progress in deep learning has recently led to great successes in many engineering applications (LeCun et al., 2015). As a prime example, convolutional neural networks (CNNs), a type of feedforward neural network, are already approaching human accuracy on visual recognition tasks including object categorization (He et al., 2015) and face recognition (Kemelmacher-Shlizerman et al., 2016). Here, we show that feedforward neural networks struggle to learn abstract visual relations that are otherwise effortlessly recognized by non-human primates (Donderi and Zelnicker, 1969; Katz and Wirght, 2006), birds (Daniel et al., 2015; Martinho III and Kacelnik, 2016), rodents (Wasserman et al., 2012) and even insects (Giurfa et al., 2001). 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 rote memorization becomes impossible (as when stimulus variability exceeds their effective capacity). The comparative success of biological neural networks in learning visual relations suggests that feedback mechanisms such as attention, working memory and perceptual grouping are the key components underlying human-level abstract visual reasoning.

Keywords: Visual Relations; Visual Reasoning; Convolutional Neural Networks; Deep Learning; Working Memory; Visual Attention

Introduction

Consider the images on 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 deep "convolutional neural network" (CNN; He et al., 2015). This is quite a remarkable feat because the network must learn to extract subtle diagnostic cues from images subject to wide variability 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 one thousand natural object categories, surpassing, for the first time, the accuracy of a human observer for the recognition of one thousand image categories on the ImageNet classification challenge.

Now, consider the image on the left side of Figure 1(b). On its face, it is quite simple compared to the images on 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 a human observer. Moreover, the ability to detect visual sameness is not unique to humans. In a striking example from Martinho III and Kacelnik (2016), newborn ducklings were shown to imprint on an abstract concept of "sameness" from birth (Figure 1(b), right panel). 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 task is difficult or even impossible for contemporary computer vision algorithms like CNNs, is known. Previous work by Fleuret et al. (2011) has shown that black-box classifiers fail on most tasks from the synthetic visual reasoning test (SVRT), a battery of twenty-three visual-relation problems, despite massive amounts of training data. More recent work has shown how CNNs, including variants of the popular LeNet (LeCun et al., 1998) and AlexNet (Krizhevsky et al., 2012) architectures, could only solve a handful of the twenty-three SVRT problems (Ellis et al., 2015; Stabinger et al., 2016). Similarly, Gülçehre and Bengio (2013), 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 were inconclusive. 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 twenty-three SVRT problems represent a diverse collection of visual relations, each problem has different image features. Thus, 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 to solve 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 (Santoro et al., 2017).

In this study, 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 twenty-three 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 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 relations problems. We find that the former 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 re-appraisal 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 currently lacking in current feedforward architectures, are necessary for abstract visual reasoning.

Experiment 1: A taxonomy of visual-relation problems

The SVRT challenge

The Synthetic Visual Reasoning Test (SVRT) is a collection of twenty-three binary classification problems in which opposing classes differ based on whether or not images obey an abstract rule (Fleuret et al., 2011). 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 twenty-three 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`.

Hyper-parameter search

We tested CNNs of three different depths (2, 4 and 6 convolutional layers) and three different convolutional receptive field sizes (2×2, 4×4 and 6×6) for a total of nine networks. All networks used pooling kernels of size 3×3, convolutional strides of 1, pooling strides of 2 and three fully connected layers. Pooling layers used ReLu activations. We trained all nine networks on each problem and selected the best-performing network for each problem. All networks were trained using the Adaptive Moment Estimation (Adam) optimizer (Kingma and Ba, 2015) with base learning rate of $\eta = 10^{-4}$. All experiments were run using TensorFlow (Abadi et al., 2016).

Figure 2. *Examples images of twenty-three SVRT problems*. For each problem, three example images, two negative and one positive, are displayed in a row. Problems are ordered and color-coded identically to Figure 3. Images in each problem respect a certain 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 respect 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 Fleuret et al. (2011).

Results

Shown in Figure 3 is a bar plot of the best-performing network accuracy for each of the twenty-three SVRT problem (sorted by accuracy). Bars are colored red or blue according to the SVRT problem descriptions given in (Fleuret et al., 2011). Problems whose descriptions have words like “same” or “identical” are colored red. These *Same-Different* (SD) problems have items that are congruent up to some transformation. *Spatial-Relation* (SR) problems, whose descriptions have phrases like “left of”, “next to” or “touching,” are colored blue. Figure 2 shows positive and negative samples for each of the corresponding twenty-three problems (also sorted by accuracy).

The resulting dichotomy across the SVRT problems is striking (Figure 3). 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 result matches earlier evidence for a visual-relation dichotomy hypothesized by Stabinger et al. (2016) which was unknown to us at the time of our own experiments.

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, larger networks yielded significantly higher accuracy on SD problems compared to 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 (Fleuret et al., 2011; Gülçehre and Bengio, 2013; Ellis et al., 2015; Stabinger et al., 2016; Santoro et al., 2017) and suggests that low accuracy cannot be simply attributed to a poor choice of hyper-parameters.

Limitations of the SVRT challenge

Though useful for surveying many types of relations, the SVRT challenge has two important limitations. First, different problems have different image features. For instance, Problem 2 (*"inside-outside"*) requires that an image contain one large object and one small object. Problem 1 (*"same-different up to translation"*), on the other hand, requires that an image contains two items identically-sized and positioned without one being contained in the other. In other cases, different problems simply require different number of objects in a single image (two items in Problem 1 vs. three in Problem 9). Overall, this leaves open the possibility that image features, not abstract relational rules, make some problems harder than others. 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. A better way to compare visual-relation problems would be instead to define various problems on the *same* set of images.

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

The PSVRT challenge

To address the limitations of SVRT, we constructed a new visual-relation benchmark consisting of two idealized problems from the dichotomy that emerged from Experiment 1 (Figure 4): *Spatial Relations* (SR) and *Same-Different* (SD). Critically, both problems in this new benchmark used the exact 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 CNN only memorizes the distinguishing features of the two image classes, then learning should be affected by the variability of these features. For example, when the image and items are large, there are simply more possible samples, which might slow down the training of a CNN trying to learn by rote memorization. In rule-based problems such as visual relations, these two behaviors can be distinguished by training and testing the same architecture on a problem instantiated over a multitude of image distributions. There is no hold-out set in this experiment. 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 generalizable representations of one set of image parameters can also learn comparably generalizable representations of another set of parameters with equal ease by taking advantage of the abstractness of the visual rule.

Methods

Our image generator produces a gray-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. These parameters allow us to quantify the number of possible images in a dataset as $\mathcal{O}(P_{n^2,k}\, 2^{km^2})$, where $P_{a,b}$ denotes the number of possible permutations of a elements from a set of size b. Our parametric construction allows a dissociation between two possible factors that may affect a problem difficulty: classification rules vs. 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 labeling 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 centers (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 2* identical items in the image, and 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°. Each image is generated by first drawing a joint class label for SD and SR from a uniform distribution over $\{Different, Same\} \times \{Horizontal, Vertical\}$. 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 k unique items are then consecutively sampled. These k items are then randomly placed in an $n \times n$ image while ensuring at least 1 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 thousand images. These samples were averaged across 10 repetitions of each condition, yielding a single, scalar measure of learnability called "average training accuracy" (ATA). In all of our experiments, accuracy either gradually increased or saturated at some fixed value. Therefore, ATA is high only when accuracy increases earlier and more rapidly throughout the course of training and if 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 uses the first strategy when learning the problem, 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, average accuracy will be consistently high over a range of image parameters. Alternatively, if the network's architecture doesn't 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, average accuracy will gradually decrease as image variability increases.

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 by 180 pixels by placing a smaller PSVRT image (if $n < 180$) at the center of a blank background of size 180 by 180 pixels. The baseline CNN was trained from scratch in each condition with 20 million training images and a batch size of 50. To examine the effect of the network size on learnability, we also repeated our experiments with a larger network control (Figure 5, purple curve) with 2 times the number of units in the convolution layers and 4 times the number of units in the fully-connected layers.

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". The ATA from a training session was determined by when this sudden rise occurred and at what accuracy it plateaued. When there was no learning event, accuracy remained at chance and ATA was 0.5.

In SR, across all image parameters over all random initializations, the learning event immediately occurred at the start of training and quickly approached 100% accuracy, producing consistently high and flat ATA curves (Figure 5, blue dotted lines). In SD, however, we found that ATA was overall significantly lower than SR even though the training images have been sampled from the same distribution. Additionally, we observed a significant straining effect from one image parameter, image size (n). Increasing image size progressively decreased ATA by making learning event progressively less likely (Figure 5, red dotted lines): the network learned SD in 7 out of 10 random initializations for the baseline parameter configuration while it only learned it in 4 out of 10 on 120×120 images. At image size 150×150 and above, the network never learned the problem. Increasing the number of items produced a slightly different straining effect. While the frequency at which learning event occurred did not change significantly, the final accuracy reached by the end of training steadily decreased from over 90% to around 80%. In contrast, increasing item size produced no visible straining effect on the CNN. Similar to SR, learnability, both in terms of the frequency of learning event as well as final accuracy, did not change significantly over the range of item sizes we considered. Using a CNN with more than twice the number of free parameters as a control did not change the qualitative trend observed in the baseline model (Figure 5, purple dotted lines).

We hypothesize that these straining effects reflect the way image size 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 2 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 sheds 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 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 data set, instead of learning the abstract "rule" per se. If a network is able to learn features that capture the visual relation at hand (a feature set to detect *any* pair of items arranged horizontally), 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 ATA for the CNNs. This suggests that the features learned by CNN are not invariant rule-detectors, but rather merely a collection of templates covering a particular distribution in the image space.

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. (2017) 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 6a). These objects correspond to feature columns coarsely sampled at all retinotopic locations from a high-level layer of a CNN (similar, in a sense, to the feature columns found in higher areas of the visual cortex, see Tanaka, 2003).

As such, 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. Santoro et al. (2017) found that an RN architecture substantially outperformed a baseline CNN on various reasoning problems. The authors emphasize that their model performed well even though it employs a 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 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 7(a) for examples of sort-of-CLEVR items). In “sort-of-CLEVR”, scenes contain up to six items, each of which has one of two shapes and six colors. The RN was trained to answer both relational questions (e.g., *“What is the shape of the object that is farthest from the gray object?”*) and non-relational questions (e.g., *“Is the red object on the top or bottom of the scene?”*). However, 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 whether 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.

Without testing the RN on this more difficult task, it is difficult to evaluate the efficacy of the model’s “unstructured” objects. If the model learns that an object is a flexible combination of any colors and shapes from its training, then it should be able to detect same-different relations among novel combinations of familiar shapes and colors. That is, it should be able to “group” these item attributes into a new object. If, on the other hand, RN object representations reflect *particular* familiar color-shape combinations, then it would not be able to transfer the concept of sameness to new combinations.

To investigate these alternatives, we trained an RN on a two-item same-different task using sort-of-CLEVR items, but leaving out certain color-shape combinations. Furthermore, to examine the efficacy of perceptual grouping on same-different problems, we introduced a novel model which forcibly groups pixels in single items into object representations.

Our new model is a “Siamese” network (Bromley et al., 1994) which processes each scene item in a separate (CNN) channel and then passes the processed items to a single classifier network. This idealized model simulates the effects attentional selection and perceptual grouping by segregating the representations of each item. Unlike an RN, whose object representations may in fact contain no item, multiple items or incomplete items, object representations in the Siamese network contain exactly one item.

Methods

Sub-experiment 3.1: Failure of 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 Wright and Kelly, 2017; 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`. This is essentially the architecture and training procedure used in the original study and we confirmed that this model was able to reproduce the results from (Santoro et al., 2017) on the sort-of-CLEVR task.

We constructed twelve different versions of the sort-of-CLEVR dataset, each one missing one of the twelve possible color $\times$ shape attribute combinations, see Figure 7(a). 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 color 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.

Sub-experiment 3.2: The need for perceptual grouping and object individuation Here, we introduce a Siamese network which processes scene items individually in separate CNN "channels" (Fig. 6b). First, we 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 (Fig. 6b). These feature maps were then concatenated before being passed to the classifier.

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.

Results

Sub-experiment 3.1: Relational transfer to novel attribute combinations From the sort-of-CLEVR transfer task, we found that the RN does not generalize on average to left-out color+shape attribute combinations (Figure 7). Since there are only 11 color+shape combinations in any given setup, the model did not need to learn to generalize across many items if it could simply memorize all combinations of “same” instances. As a result, the RN learned orders of magnitude faster than the CNNs in Experiment 2. However, while the average training accuracy curve (solid red) 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).

Sub-experiment 3.2: The need for perceptual grouping and object individuation We ran the Siamese model on the PSVRT tasks, again measuring ATA. The ATA curves for the Siamese network were strikingly different from that of the CNN in Experiment 2 (Figure 8). Barely any straining effect was observed on the SD task, and the model learned within 5M examples across all image size parameters. 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.

Indeed, in informal experiments (data not shown) we found that very shallow Siamese networks (e.g. with one convolutional layer) could still learn SD much faster than baseline CNNs. These results indicate that object individuation makes same-different problems trivially easy. 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 using the mechanisms explored in the Discussion section.

Discussion

Recent progress in computational vision has been significant. Modern deep learning architectures can discriminate between one thousand object categories (He et al., 2015) or identify faces among millions of distractors (Kemelmacher-Shlizerman et al., 2016) at a level approaching – and possibly 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 Serre, 2015; Kriegeskorte, 2015; for reviews). Further, CNNs have been shown to account well for monkey inferotemporal data (Yamins et al., 2014) and human lateral occipital data (Khaligh-Razavi and Kriegeskorte, 2014; Guclu and van Gerven, 2015). In addition, deep networks have been shown to be consistent with a number of human behaviors including rapid visual categorization (Eberhardt et al., 2016), image memorability (Dubey et al., 2015), typicality (Lake et al., 2015b) as well as similarity (Peterson et al., 2016) and shape sensitivity (Kubilius et al., 2016) judgments.

At the same time, there is a growing body of literature highlighting key dissimilarities between current deep network models and various aspects of visual cognition. One prominent example is adversarial perturbation (Goodfellow et al., 2015), structured image distortions that asymmetrically affects CNNs compared to human participants. 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 are effortless for 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 (Lake et al., 2015a; Saleh et al., 2016; Erdogan and Jacobs, 2017). Direct evidence for qualitatively different visual strategies used by humans and CNNs was shown in (Ullman et al., 2016; Linsley et al., 2017).

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 been long acknowledged by cognitive scientists (Fodor and Pylyshyn, 1988; Marcus, 2001). However, this limitation seems to have been overlooked by current computer vision scientists.

Compared to the feedforward networks in this study, biological visual systems excel at detecting relations. Fleuret et al. (2011) 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 6 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."* In contrast, the best performing CNN model for this problem could not get significantly above chance from one million training examples.

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 (Donderi and Zelnicker, 1969; Katz and Wirght, 2006), birds (Daniel et al., 2015; Martinho III and Kacelnik, 2016), rodents (Wasserman et al., 2012) and even insects (Giurfa et al., 2001) can be trained to recognize abstract relations between training objects and then transfer this knowledge to novel objects. Contrast the behavior of these ducklings 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 the neural substrate of visual-relation detection 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 (Serre, 2016). However, psychophysical evidence suggests that this feedforward sweep is too spatially coarse to localize objects even when they can be recognized (Evans and Treisman, 2005). The implication is that object localization in clutter requires attention (Zhang et al., 2011).

It is difficult to imagine how one could recognize a relation between two objects without spatial information. Indeed, converging evidence (Logan, 1994; Moore et al., 1994; Rosielle et al., 2002; Holcombe et al., 2011; Franconeri et al., 2012; van der Ham et al., 2012) 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 (Kroger et al., 2002; Golde et al., 2010; Clevenger and Hummel, 2014; Brady and Alvarez, 2015). In particular, imaging studies (Kroger et al., 2002; Golde et al., 2010) 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 (Franconeri et al., 2012) 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 associated with feedforward neural networks.

Humans can easily detect when two objects are the same up to some transformation (Shepard and Metzler, 1971) or when objects exist in a given spatial relation (Fleuret et al., 2011; Franconeri et al., 2012). More generally, humans can effortlessly construct an unbounded set of structured descriptions about their visual world (Geman et al., 2015). Given the vast superiority of humans over modern computers in their ability to detect visual relations, we see the exploration of attentional and mnemonic mechanisms as an important step in our computational understanding of visual reasoning.

Acknowledgments

The authors would like to thank Drs. Drew Linsley and Sven Eberhardt for their advice, along with Dan Shiebler for earlier work. This research was supported by NSF early career award [grant number IIS-1252951] and DARPA young faculty award [grant number YFA N66001-14-1-4037]. Additional support was provided by the Center for Computation and Visualization (CCV) at Brown University. MR is supported by an NSF Graduate Research Fellowship.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, Berkeley, CA, USA. USENIX Association.

Brady, T. F. and Alvarez, G. A. (2015). Contextual effects in visual working memory reveal hierarchically structured memory representations. *J. Vis.*, 15:1–69.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1994). Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, pages 737–744.

Clevenger, P. E. and Hummel, J. E. (2014). Working memory for relations among objects. *Attention, Perception, Psychophys.*, 76:1933–1953.

Daniel, T. A., Wright, A. A., and Katz, J. S. (2015). Abstract-concept learning of difference in pigeons. *Anim. Cogn.*, 18(4):831–837.

Donderi, D. and Zelnicker, D. (1969). Parallel processing in visual same-different. *Percept. Psychophys.*, 5(4):197–200.

Dubey, R., Peterson, J., Khosla, A., Yang, M.-H., and Ghanem, B. (2015). What makes an object memorable? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1089–1097.

Eberhardt, S., Cader, J. G., and Serre, T. (2016). How deep is the feature analysis underlying rapid visual categorization? In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 1100–1108. Curran Associates, Inc.

Ellis, K., Solar-lezama, A., and Tenenbaum, J. B. (2015). Unsupervised Learning by Program Synthesis. *Neural Information Processing Systems*, pages 1–9.

Erdogan, G. and Jacobs, R. A. (2017). Visual shape perception as bayesian inference of 3D object-centered shape representations. *Psychol. Rev.*, 124(6):740–761.

Evans, K. K. and Treisman, A. (2005). Perception of objects in natural scenes: is it really attention free? *J. Exp. Psychol. Hum. Percept. Perform.*, 31(6):1476–1492.

Fleuret, F., Li, T., Dubout, C., Wampler, E. K., Yantis, S., and Geman, D. (2011). Comparing machines and humans on a visual categorization test. *Proc. Natl. Acad. Sci. U. S. A.*, 108(43):17621–5.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., and Kahn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition*, 122(2):210–227.

Geman, D., Geman, S., Hallonquist, N., and Younes, L. (2015). Visual Turing test for computer vision systems. *Proc. Natl. Acad. Sci. U. S. A.*, 112(12):3618–3623.

Giurfa, M., Zhang, S., Jenett, A., Menzel, R., and Srinivasan, M. V. (2001). The concepts of 'sameness' and 'difference' in an insect. *Nature*, 410(6831):930–933.

Golde, M., von Cramon, D. Y., and Schubotz, R. I. (2010). Differential role of anterior prefrontal and premotor cortex in the processing of relational information. *Neuroimage*, 49(3):2890–2900.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *Iternational Conference on Learning Representations*.

Guclu, U. and van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014.

Gülçehre, Ç. and Bengio, Y. (2013). Knowledge Matters : Importance of Prior Information for Optimization. *arXiv Prepr. arXiv1301.4083*, pages 1–12.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving Deep into Rectifiers: Surpassing

Human-Level Performance on ImageNet Classification. *arXiv Prepr. arXiv1502.01852*, pages 1–11.

Holcombe, A. O., Linares, D., and Vaziri-Pashkam, M. (2011). Perceiving spatial relations via attentional tracking and shifting. *Curr. Biol.*, 21(13):1135–1139.

Katz, J. S. and Wirght, A. A. (2006). Same/different abstract-concept learning by pigeons. *J. Exp. Psychol. Anim. Behav. Process.*, 32(1):80–86.

Kemelmacher-Shlizerman, I., Seitz, S. M., Miller, D., and Brossard, E. (2016). The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.*, 10(11):e1003915.

Kingma, D. P. and Ba, J. L. (2015). Adam: a method for stochastic optimization. In *International Conference on Learning Representations*.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annu Rev Vis Sci*, 1:417–446.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.*

Kroger, J. K., Sabb, F. W., Fales, C. L., Bookheimer, S. Y., Cohen, M. S., and Holyoak, K. J. (2002).

Recruitment of Anterior Dorsolateral Prefrontal Cortex in Human Reasoning: a Parametric Study of Relational Complexity. *Cereb. Cortex*, 12(5):477–485.

Kubilius, J., Bracci, S., and Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput. Biol.*, 12(4):e1004896.

Lake, B., Salakhutdinov, R., and Tenenbaum, J. (2015a). Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Lake, B. M., Zaremba, W., Fergus, R., and Gureckis, T. M. (2015b). Deep neural networks predict category typicality ratings for images. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2323.

Linsley, D., Eberhardt, S., Sharma, T., Gupta, P., and Serre, T. (2017). What are the visual features underlying human versus machine vision? In *IEEE ICCV Workshop on the Mutual Benefit of Cognitive and Computer Vision*.

Logan, G. D. (1994). Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, 20(5):1015–1036.

Marcus, G. (2001). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. MIT Press, Cambridge, MA.

Martinho III, A. and Kacelnik, A. (2016). Ducklings imprint on the relational concept of "same or different". *Science*, 353(6296):286–288.

Moore, C. M., Elsinger, C. L., and Lleras, A. (1994). Visual attention and the apprehension of spatial relations: The case of depth. *J. Exp. Psychol. Hum. Percept. Perform.*, 20(5):1015–1036.

Peterson, J., Abbott, J., and Griffiths, T. (2016). Adapting deep network features to capture psychological representations. In Grodner, D., Mirman, D., Papafragou, A., and Trueswel, J., editors, *38th annual conference of the cognitive science society*, pages 2363–2368.

Rosielle, L. J., Crabb, B. T., and Cooper, E. E. (2002). Attentional coding of categorical relations in scene perception: evidence from the flicker paradigm. *Psychon. Bull. Rev.*, 9(2):319–26.

Saleh, B., Elgammal, A., and Feldman, J. (2016). The role of typicality in object classification: Improving the generalization capacity of convolutional neural networks.

Santoro, A., Raposo, D., Barrett, D. G. T., Malinowski, M., Pascanu, R., Battaglia, P., and Lillicrap, T. (2017). A simple neural network module for relational reasoning. *arXiv Prepr. arXiv1706.01427*.

Serre, T. (2015). Hierarchical models of the visual system. In Jaeger, D. and Jung, R., editors, *Encyclopedia of Computational Neuroscience*, pages 1309–1318. Springer New York.

Serre, T. (2016). Models of visual categorization. *Wiley Interdiscip. Rev. Cogn. Sci.*, 7(3):197–213.

Shepard, R. N. and Metzler, J. (1971). Mental Rotation of Three-Dimensional Objects. *Science*, 171(3972):701–703.

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

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(1):90–99.

Ullman, S., Assif, L., Fetaya, E., and Harari, D. (2016). Atoms of recognition in human and computer vision. *Proc. Natl. Acad. Sci. U. S. A.*, 113(10):2744–2749.

van der Ham, I. J. M., Duijndam, M. J. A., Raemaekers, M., van Wezel, R. J. A., Oleksiak, A., and Postma, A. (2012). Retinotopic mapping of categorical and coordinate spatial relation processing in early visual cortex. *PLoS One*, 7(6):1–8.

Wasserman, E. A., Castro, L., and Freeman, J. H. (2012). Same-different categorization in rats. *Learn. Mem.*, 19(4):142–145.

Wright, A. A. and Kelly, D. M. (2017). Comparative approaches to same/different abstract concept-learning. *Learn. Behav.*, 45:323–324.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.*, 111(23):8619–8624.

Zhang, Y., Meyers, E. M., Bichot, N. P., Serre, T., Poggio, T., and Desimone, R. (2011). Object decoding with attention in inferior temporal cortex. *Proc. Natl. Acad. Sci. U. S. A.*, 108(21):8850–8855.

Figures/fig1_horizontal.png

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. *(b)* In addition to categorizing visual objects, humans can also perform comparison between objects and determine if they are identical up to a rotation (left). The ability to recognize "sameness" is also observed in other species in the animal kingdom such as birds (right). The geometric figures are adapted from (Shepard and Metzler, 1971), and the image with a duckling is taken with permission from (Martinho III and Kacelnik, 2016).

Figures/svrt_examples.pdf

Figures/svrt_bars.png

Figure 3. *SVRT results*. Multiple CNNs with different combinations of hyper-parameters were trained on each of the twenty-three SVRT problems. Shown are the ranked accuracies of the best-performing network 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.

Figures/PSVRT/examples/basic/PSVRT_examples.png

Figure 4. *The PSVRT challenge. (Left)* Four images show the joint categories of SD (grouped by columns) and SR (grouped by rows) tasks. Our image generator is designed such that each image can be used to pose both problems by simply labeling 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: $m = 4$, $n = 60$, $k = 2$. *(Right)* Six example images show different choices of image parameters used in our experiment: item size, number of items and image size. All images shown here belong to *Same* and *Vertical* categories. When more than 2 items are used, SD category label is determined by whether there are at least two identical items in the image. 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°.

Figures/PSVRT/results/AUC/psvrt_res_horizontal.png

Figure 5. *Average Training Accuracy (ATA) curves over PSVRT image parameters.* ATA denotes the average value of accuracy in each experimental condition measured over the course of 20 million training images and over 10 random initializations. Three curves – SD (red), SD with a large CNN control, (purple) and SR (blue) – are plotted. The three figures display average training accuracy curves over each of three image variability parameters: item size, image size and number of items.

Figures/rn_diag.png

(a)

Figures/siamese_diag.png

(b)

Figure 6. *A comparison between a relational network and the proposed Siamese architecture.* (a) A relational network (panel (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 neuron, 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 automates the attentional and perceptual grouping processes suspected to underlie biological visual reasoning (see Discussion).

Figures/rn_item_details.png Figures/rn_lcs.png

(a) (b)

Figure 7. *(a) Sample items used during training and testing in Experiment 3.* We trained relational networks on twelve two-item same-different data sets each missing one color-shape combination from sort-of-CLEVR (2 shapes $\times$ 6 colors). Then, we tested the model on the left-out combination. The top and middle rows of panel *(a)* show two possible pairs of item 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 data sets missing one color-shape combination. The red curve shows the training accuracy. The blue dashed line shows the accuracy on validation data with the left-out items.

Figures/mauc_img_size.pdf

Figure 8. *Average Training Accuracy (ATA) curves for CNN and Siamese model on a two-item same-different (SD) task.* The CNN's ATA curve (red) is taken from Experiment 2. The Siamese network's ATA curve (green) indicates almost no straining. The network learns equally well on large images, for which there is great positional variety of items, as it does on small images, for which there is much less variety.