NOT-SO-CLEVR: VISUAL RELATIONS STRAIN FEEDFORWARD NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

The robust and efficient recognition of visual-relation in images is a hallmark of biological vision. Here, we argue that, despite recent progresses in visual recognition, modern machine vision algorithms are severely limited in their ability to learn visual relations. Through controlled experiments, we demonstrate that visual-relation problems strain convolutional neural networks (CNNs). The networks eventually break altogether when rote memorization becomes impossible such as when the intra-class variability exceeds their capacity. We further show that another class of feedforward networks called relational networks (RNs) which were shown to successfully solve seemingly challenging visual question answering (VQA) challenges on the CLEVR datasets, suffer the same limitations. Motivated by the comparable success of biological vision, we argue that the incorporation of feedback mechanisms including working memory and attention will constitute a necessary step towards building machines that are capable of abstract visual reasoning.

1 INTRODUCTION

Consider the two images in Fig. 1. The image on the left was correctly classified as a flute by a deep convolutional neural network (CNN) (He et al., 2015). This is quite a remarkable feat for such a complicated image, which includes distractors that partially occlude the object of interest. After the network was trained on millions of photographs, this and many other images were accurately categorized into one thousand natural object categories, surpassing, for the first time, the accuracy of a human observer on the ImageNet classification challenge.

Now, consider the image on the right. On its face, it is quite simple compared to the image on the left. It is just a binary image containing two curves. Further, it has a rather distinguishing property, at least to the human eye: both curves are the same. The relation between the two items in this simple scene is rather intuitive and immediately obvious to a human observer. Yet, the CNN failed to learn this relation even after seeing millions of training examples.

Why is it that a CNN can accurately detect the flute in Fig. 1a while struggling to recognize the simple relation depicted in Fig. 1b?

(a) (b)

Figure 1: *Two images:* The image in panel *(a)* can be classified with high confidence as containing a flute by contemporary computer vision algorithms. However, these same algorithms struggle to learn the concept of "sameness" as exemplified by the image with the two curves shown in panel *(b)*. The image in panel *(b)* is sampled from the SVRT challenge (Fleuret et al., 2011).

That such task is difficult, and even sometimes impossible for contemporary computer vision algorithms including CNNs, is known (Fleuret et al., 2011; Gülçehre and Bengio, 2013; Ellis et al., 2015) but has, so far, been overlooked. To make matters worse, the issue has been overshadowed by the recent successes obtained with a novel class of neural networks called relational networks (RNs) on seemingly challenging visual question answering (VQA) benchmarks. However, RNs have so far only been tested using toy datasets including the sort-of-CLEVR dataset which depicts combinations of items of only a handful of colors and shapes (Santoro et al., 2017). As we will show, RNs suffer the same limitations as CNNs for a same-different task such as the one shown in Fig. 1b.

This failure of modern computer vision algorithms is all the more striking given the widespread ability to recognize visual relations across the animal kingdom from human and non-human primates (Donderi and Zelnicker, 1969; Katz and Wirght, 2006) to rodents (Wasserman et al., 2012), birds (Daniel et al., 2015; Martinho III and Kacelnik, 2016) and even insects (Giurfa et al., 2001). Understanding the failures of existing models is a critical step on the path to creating genuine visual reasoning machines. Yet, to our knowledge, there has not been any systematic exploration of the limits of contemporary machine learning algorithms on relational reasoning problems.

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 by Ellis et al. (2015) has shown that two specific CNN architectures, i.e., LeNet (LeCun et al., 1998) and AlexNet (Krizhevsky et al., 2012), could only solve at most six of the twenty-three SVRT problems. 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 (MLP) on this task by providing carefully engineered training schedules. This leaves open the possibility that the failure of feedforward neural networks to solve various visual-relation problems reflects a poor choice of hyper-parameters rather than a systematic failure of the entire class of models.

Here, we propose to systematically probe the limits of CNNs and other state-of-the-art visual reasoning networks (RNs) on visual-relation tasks. Through a series of controlled experiments, we demonstrate that visual-relation tasks strain CNNs and that these limitations are not alleviated in RNs, which were specifically designed to tackle visual-relation problems. A brief review of the biological vision literature suggests that two key brain mechanisms, working memory and attention, underlie primates' ability to reason about visual-relation. We argue that these mechanisms and possibly other feedback mechanisms are needed to extend current computer models to efficiently learn to solve complex visual reasoning tasks.

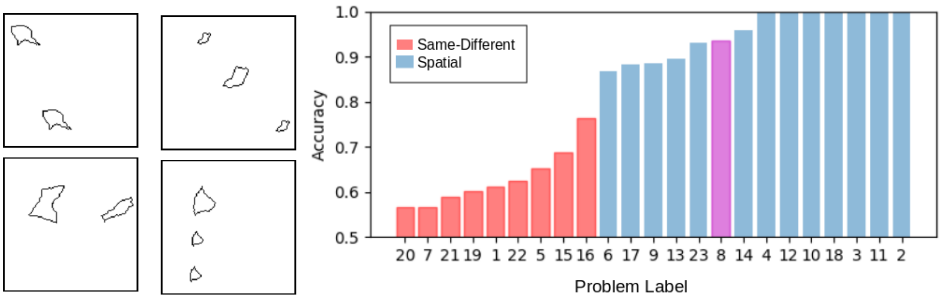
Our contributions are threefold: (i) We perform the first systematic performance analysis of CNN architectures on each of the twenty-three SVRT problems. This yields a novel dichotomy of visual-relation problems in terms of same-different vs. spatial-relation problems. (ii) We describe a novel, controlled, visual-relation challenge which convincingly shows that CNNs solve same-different tasks via rote memorization. (iii) We show that a simple modification of the sort-of-CLEVR challenge similarly breaks state-of-the-art relational network architectures.

Overall, we hope to move the field forward by prompting the community to reconsider existing visual question answering challenges and turn to neuroscience for inspiration to help with the design of visual reasoning architectures.

2 EXPERIMENT 1: SVRT

The SVRT challenge: This challenge is a collection of twenty-three binary classification problems in which opposing classes differ based on whether their stimuli 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, whereas negative examples do not (Fig. 2a). In problem 9, positive examples have three items, the largest of which is in between the two smaller ones (Fig. 2b). All stimuli depict simple, closed, black curves on a white background.

High-throughput screening approach: 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. Convolutional layers



(a) (b) (c)

Figure 2: *The Synthetic Visual Reasoning Test*. Nine CNNs corresponding to 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 for each problem. The x-axis indicates the arbitrary problem label provided in (Fleuret et al., 2011). CNNs from this high-throughput analysis were found to produce uniformly lower accuracies on same-different problems (red bars) than on spatial-relation problems (blue bars). The single purple bar corresponds to a problem which required detecting both a same-different relation and a spatial relation simultaneously.

were followed by max-pooling layers with size 3×3 kernels and a stride of 2. Pooling layers used ReLu activations. For each of the twenty-three problems, we generated 2 million examples split evenly into training and test sets using code publicly provided by Fleuret et al. (2011) at `http://www.idiap.ch/~fleuret/svrt/`. We trained all nine networks on each problem for a total of $n = 207$ conditions. All networks were trained using the ADAM optimizer with base learning rate of $\eta = 10^{-4}$.

Results: The accuracy of the best networks obtained for each problem individually (across all networks) is shown in Fig. 2c. After best-case accuracies for the twenty-three problems was obtained, we sorted the problems by accuracy and then colored the bars red or blue according to the SVRT problem descriptions provided by (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 (e.g., Problem 1 in Fig. 2a). *Spatial-Relation* (SR) problems, whose descriptions have phrases like "left of", "next to" or "touching," are colored blue.

The resulting dichotomy across the SVRT problems is striking. Evident from Fig. 2c is the fact that CNNs fare much 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 and 21) resulted in accuracies not substantially above chance. From this analysis, it appears as if SD tasks pose a particularly difficult challenge to CNNs. Additionally, our hyperparameter search revealed that the SR problems are generally equally well-learned across all network configurations with less than 10% difference in final accuracy between the worst-case and the best-case. On the other hand, larger networks generally yielded significantly higher accuracies on SD problems than smaller ones. If only the results from a single architecture had been reported, the visible dichotomy would have been stronger.

Limitations of the SVRT challenge: While the SVRT challenge is useful for surveying the effectiveness of an algorithm on a diverse range of visual-relation, it has two important limitations. First, the twenty-three problems used in the challenge constitute a somewhat arbitrary sample from a very large set of all conceivable visual-relations. While there are some obvious connections between different problems (e.g., *"same-different up to translation"* in Problem 1 and *"same-different up to scale"* in Problem 19), a direct comparison between most problems is generally hard because they often assume different image structures, each requiring unique image generation methods resulting in different image distributions. For example, Problem 2 (*"inside-outside"*) requires that an image contains one large object and one small object. This necessary configuration naturally conflicts with other problems such as Problem 1 (*"same-different up to translation"*) where two items must be comparably-sized and positioned without one being contained in the other. In other cases, problems simply require different number of objects in a single image (two items in Problem 1 vs. three in

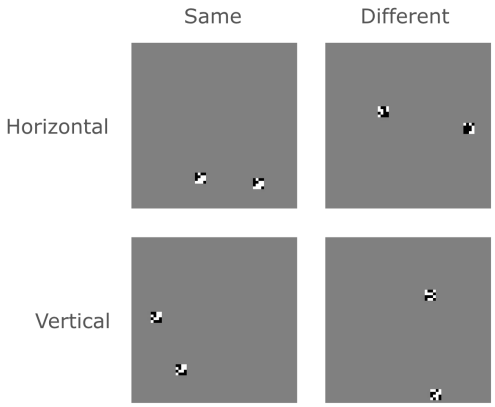


Figure 3: *Sample PSVRT images*. Four images are shown representing the four joint categories of SD (grouped by columns) and SR (grouped by rows). An image is considered *Same* or *Different* depending on whether it contains identical (left column) or dissimilar (right column) square bit patterns. An image is considered *Horizontal* (top row) or *Vertical* (bottom row) depending on whether the orientation of the displacement between the items is greater than $45°$. The images were generated with the baseline image parameters: $m = 4$, $n = 60$, $k = 2$.

Problem 9). Instead, a better way to compare visual-relation problems would be to define various problems on the *same* set of images.

Second, using simple, closed curves as items in SVRT images makes it difficult to quantify and control image variability as a function of the image generation parameters. While closed curves are perceptually interesting objects, highly procedural nature of the parameters used to generate them prevents robust quantification of image variability and its impact 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 results because of the particular choice of image parameters unrelated to the rule.

3 EXPERIMENT 2: PSVRT

The PSVRT challenge: To address the issues associated with SVRT, we constructed a new dataset consisting of two idealized problems from the dichotomy that emerged from experiment 1 (Fig. 3): *Spatial Relations* (SR) and *Same-Different* (SD). 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 identical items. As long as we limit the problems to these two simple rules, the same image dataset can be used in both problems by simply labeling each image according to different rules (Fig. 3).

Our image generator produces a gray-scale image by placing square-shaped binary bit patterns (consisting of values 1 and -1) on a blank background (consisting of value 0). Specifically, the generator uses three parameters to control image variability: item size, image size and number of items in a single image. Item size (m) refers to side-length of the square bit patterns which controls image variability at item level. Image size (n) refers to side-length of the input image. It thus controls image variability by setting the spatial extent of the placement of individual items. Lastly, number of items (k) controls both item and spatial variability because adding one more item in the image increase the total number of possible images by a factor equal to the number of different bit patterns in the new item times the number of positions to place it on.

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 $45°$. These parameters allowed 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. To highlight the parametric nature of the image samples, we call this test *Parametric* SVRT, or PSVRT.

Each image is generated by first drawing a joint class label for SD and SR from a uniform distribution over $\{\textit{Different}, \textit{Same}\} \times \{\textit{Horizontal}, \textit{Vertical}\}$. The first item is sampled from a uniform distribution in $\{-1, 1\}^{m \times m}$. Then, if the sampled SD label is *Same*, identical copies of random number between 1 and $k - 1$ are made. If the sampled SD label is *Different*, no identical copies are made. The rest of k unique items are consecutively sampled. These k items are then randomly

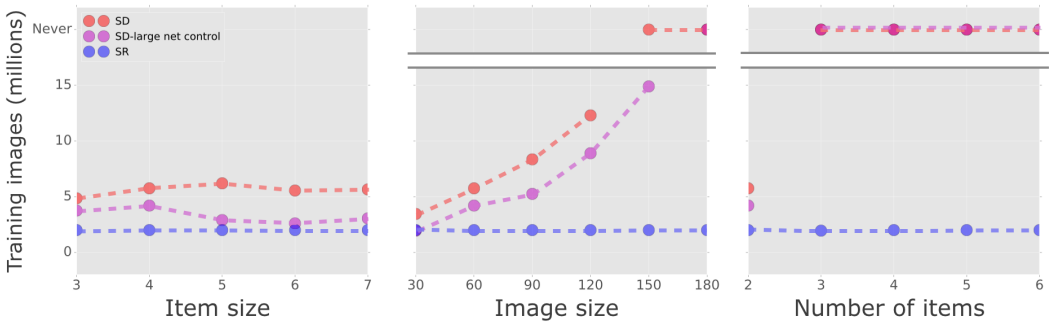


Figure 4: *Training-to-acquisition (TTA) curves over PSVRT image parameters.* TTA denotes the number of training examples needed for a CNN to reach 95% validation accuracy. Training contained 20 million images, and if accuracy never reached 95% in all of 10 random initializations we consider the problem "never" learned. The figures only show the minimum TTAs out of 10 random initializations in each condition. Three curves – SD (red), SD with a large CNN, (purple) and SR (blue) – are plotted. Three figures each display TTA curves over each of three image variability parameters: item size *(a)*, image size *(b)* and number of items *(c)*.

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.

Experimental parameters: One way of understanding the relative difficulties of different visual-relation problems is to examine whether a network can learn features that generalize to different levels of image variability. Generalization of a *trained architecture* is typically examined by comparing accuracies obtained from training and validation datasets. However, a network architecture that can learn and generalize well on a particular instance of a problem may still prove unable to learn the same problem with greater image variability. This, if true, means that the architecture can only construct a feature set that can fit well to a particular data distribution, but not to the visual-relation itself. Therefore, the effectiveness of an architecture to learn visual-relation problems should be measured in terms of generalization over multiple variants of the same problem, not over multiple splits of the same dataset.

For this, we first established a baseline architecture and image parameters at which a problem is learned with high accuracy (95%). We then observed how quickly learning deteriorates as image variability increases. If a network is, by construction, able to learn features that generalize well while capturing the visual relationship (e.g., a feature set to detect *any* pair of items arranged horizontally), then these features should be minimally sensitive to the image variations that are irrelevant to the visual relation. This means that, in addition to the learned feature set generalizing well to validation data, the same architecture should also be able to learn the same problem with increasing levels of image variability by constructing a similar feature set. However, if a network is architecturally unable to learn a generalizable feature set (e.g., it can only learn by rote memorization of the examples obeying a relation), then its ability to learn the problem will be constrained to small levels of image variability and quickly deteriorate beyond that range.

The baseline convolutional network we used has four convolution and pool layers and four fully-connected layers. The first convolution layer has 16 kernels of size 4×4, followed by 32, 64 and 128 kernels of size 2×2 in the subsequent convolution layers. Four pool layers are interleaved after each convolution step with a kernel of size 3×3 and with strides of size 2×2. The fully-connected layers have 256 units in each layer. We used dropout in the last fully-connected layer with probability 0.5. We used ADAM optimizer with base learning rate 1×10^{-4}. To examine the effect of the network size on learnability, we also repeated our experiments with a larger network control (Fig. 4, 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. The baseline image parameters used were $n = 60, n = 4, k = 2$. We varied each image parameter separately to examine its effect on learnability. n was varied between 30 and 180; n between 3 and 7; k between 2 and 6. A CNN was trained with each condition over 20 million training images with batch size of 50. We only report the best-case result for each experimental condition out of 10 random initializations.

Results: In all conditions, we found a strong dichotomy in the observed learning curves. In conditions where learning occurs (accuracy reached 95%), training accuracy suddenly shoots up from chance-level and then gradually approaches 100% accuracy. Although we have observed some variation in the point at which shoot-up takes place and the rate at which 95% accuracy is reached, the training runs that exhibited shoot-up also almost invariably reached 95% accuracy within 20 million training images. On the other hand, when learning doesn't happen shoot-up also almost never took place over the entire length of training. Thus, the final accuracy over different experimental conditions exhibited a strong bi-modality – chance-level or close to 100%. Thus, it is sufficient to consider training-to-acquisition (TTA), the number of examples a network has been trained on until it reached 95% accuracy, as a proxy for learnability. In Fig. 4, we report minimum TTA in each experimental condition over 10 random initializations.

In SR, we found no straining effect in all problem parameters over all random initializations. Shoot-up took place immediately after training begins and soon reached 95% thereafter. In SD, learning was also immediate and robust similar to SR in the baseline condition. However, we found a significant straining effect from two image parameters, image size (n) and number of items (k). For example, increasing image size progressively delays TTA *while also making shoot-up less likely*. As a result, the network learned SD in 7 out of 10 random initializations in the baseline condition while it only learned it in 4 out of 10 on 120×120 images. At image size of 150×150 and above, the network never exhibited shoot-up and thus never learned the problem. Increasing the number of items produced an even stronger straining effect. In fact, the network never learned the problem when there were 3 or more items in an image.

We hypothesize that these straining effects reflect the way these two parameters contribute to image variability. As we have shown above, image variability is an exponential function of image size as the base and number of items as the exponent. Thus, increasing image size while keeping the number of items as 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. The straining effect was equally strong between two CNNs with 4 times difference in network width, only with a constant offset only in the TTA curve over image sizes; the transition to the problem being essentially impossible was only delayed by one step in image size.

In contrast, increasing item size produced no visible straining effect on CNN. Much similar to SR, learnability is preserved and stable over the range of item sizes we considered. While we realize that it is possible to construct feedforward feature detectors that can generalize to coordinated item variability (the bit patterns themselves can vary arbitrarily but they should always vary together) such as "subtraction templates" with distinct excitatory and inhibitory regions in a particular spatial arrangement, we have yet to find more direct supporting evidence for such features actually being learned via training in SD.

4 EXPERIMENT 3: RELATIONAL NETWORKS

The Relational Network: Recently, Santoro et al. (2017) proposed the relational network (RN), an architecture explicitly designed to detect visual relations, and tested it on several VQA tasks. This simple feedforward network sits on top of a CNN and learns a map from pairs of high-level CNN feature vectors to the answers of relational questions. Relational questions are either provided to the model as natural language which is then processed by a long short-term memory (LSTM) or simply as hardcoded binary strings. The entire system (CNN+LSTM+RN) can be trained end-to-end. The approach was found to substantially outperform a baseline CNN on various visual reasoning problems.

Sort-of-CLEVR: In particular, an RN was able beat a CNN on "sort-of-CLEVR", a VQA task using images with with simple 2D items (Fig. 5a). Scenes had up to six items, each of which had one of two shapes (circle or square) and six colors (yellow, green, orange, blue, red, gray). 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 blue object on the top or bottom of the scene?"*).

Limitations of sort-of-CLEVR: The sort-of-CLEVR tasks suffers from two key shortcomings. First, although solving the task requires comparing the attributes of cued items, it does not necessitate learning the concept of sameness per se (*"Are any two items the same in this scene?"*). Second, there are only twelve possible items (2 shapes $\times$ 6 colors). Low item variability encourages the RN to solve relational problems using rote memorization of all possible item configurations. In order to understand how RNs perform when these handicaps are removed, we trained the model on both a two-item sort-of-CLEVR Same-Different task and on PSVRT stimuli.

Architecture details: We used software for relational networks publicly available at `https://github.com/gitlimlab/Relation-Network-Tensorflow`. The convolutional network component of the model had four convolutional layers with kernel sizes of 5×5 with ReLu activations but no intermittent pooling. The stride was set to 3 in the first two convolutional layers and 2 in the second two. There were 24 features per layer. The RN part of the system comprised a 4-layer MLP with 256 units per layer followed by a 3-layer MLP with 256 units per layer. All fully connected layers in the system except for the last one used ReLu activations. The penultimate layer was trained with 50% dropout. The output of the final layer was passed through a softmax function and the whole system was trained with a cross-entropy loss with an ADAM optimizer with base learning rate $2.5\times^{-4}$. Weights were initialized using Xavier initialization. This is essentially the exact architecture and training procedure used by the original authors (though they did not provide kernel sizes or strides) and we confirmed that this model was able to reproduce the results from (Santoro et al., 2017) on the sort-of-CLEVR task.

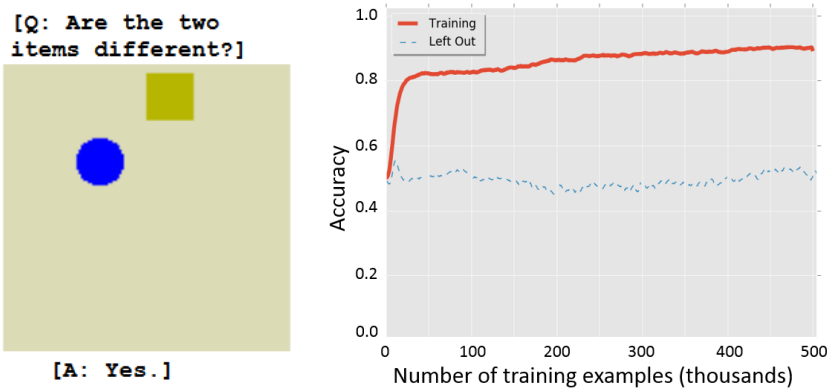
Results: We constructed twelve different versions of the sort-of-CLEVR dataset, each one missing one of the twelve possible color+shape combinations. Images in each dataset only depicted two items. Half of the time, these items were the same (same color and same shape). For each dataset, we trained our CNN+RN architecture to detect the possible sameness of the two scene items while measuring validation accuracy on the left-out images. Learning terminated when the model reached 95% training accuracy. We then averaged training accuracy and validation accuracy across all of the left-out conditions.

We found that the CNN+RN generalizes poorly to left-out color+shape combinations on the sort-of-CLEVR task (Fig. 5b). Since there are only 11 color+shape combinations in any given setup, the model does not need to learn to generalize across many items and therefore learns orders of magnitude faster than CNNS on PSVRT stimuli. However, while the average training accuracy curve (solid red) rises rapidly to around 90%, the average validation accuracy remains at chance. In other words, on average, there is no transfer of same-different ability to the left out condition, even though the attributes from that condition (e.g., blue square) were represented in the training set, just not in that combination (e.g., blue circle and red square).

Next, we replaced the simple shapes of sort-of-CLEVR with PSVRT bit patterns. As in Experiment 2, we varied image size from 30 to 180 pixels in increments of 30 while measuring learnability. We trained on 40M images. We repeated training over three different runs to make sure results were stable. We found that the combined CNN+RN behaves essentially like a vanilla CNN . After a long period at chance-level performance over several million images, the CNN+RN leaps to greater than 95% accuracy as long as the spatial extent is 120 or below. For spatial extents of 150 and 180, the system did not learn. We speculate that this cutoff point corresponds to the representational capacity of our particular RN architecture. Although we demonstrated this capacity was sufficient to solve the original sort-of-CLEVR task, it is clearly not enough for some same-different tasks on PSVRT.

5 DISCUSSION

Our results indicate that visual-relation problems can quickly exceed the representational capacity of CNNs. While learning templates for individual objects appears to be quite tractable for today's deep networks, learning templates for *arrangements* of objects become rapidly intractable because of the combinatorial explosion in the number of templates needed. That stimuli with a combinatorial structure are difficult to represent with feedforward networks has been long acknowledged by cognitive scientists at least as early as (Fodor and Pylyshyn, 1988). However, this limitation seems to have been somehow overlooked by current computer vision scientists.



(a) (b)

Figure 5: *(a)* An example of two-item same-different problem posed on a sort-of-CLEVR image. *(b)* Accuracy curves of an RN while being trained in two-item same-different problem on sort-of-CLEVR dataset with one of twelve (2 shapes $\times$ 6 colors) item types left out. The red curve shows the accuracy on validation data generated using the same set of items used for training. The blue dashed line shows the accuracy on validation data generated using the left-out items.

Biological visual systems excel at detecting visual-relations, and there is substantial evidence that this ability may require 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, there exists psychophysical evidence suggesting that this feedforward sweep is spatially too 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 the spatial relation between two objects without spatial information. Indeed, converging neuroscience 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 participants are able to detect the presence of the individual objects pre-attentively, presumably in a single feedforward sweep.

Another brain mechanism that has been 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 premotor cortices when participants solve Raven's progressive matrices which requires both spatial and same-different reasoning.

What is the computational role of attention and 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 representation built "on-the-fly" circumvents the combinatorial explosion associated with the storage of templates for all possible relations, helping to prevent the capacity overload associated with feedforward neural networks learning visual-relations.

Humans can easily detect when two objects are the same up to some transformation (Shepard and Metzler, 1971), 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 the visual world around them (Geman et al., 2015). Given the vast superiority of humans over modern computers in their ability to detect visual-relations, we see the incorporation of attentional and mnemonic mechanisms as an important step in the creation of visual reasoning machines that rival human capabilities.

REFERENCES

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

Clevenger, P. E. and Hummel, J. E. (2014). Working memory for relations among objects. *Attention, Perception, Psychophys.*, 76(December 2013):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. O. N. C. and Zelnicker, D. (1969). Parallel processing in visual same-different. *Percept. Psychophys.*, 5(4):197–200.

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

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.

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.

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.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.*, pages 1–9.

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.

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

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

Martinho III, A. and Kacelnik, A. (2016). Ducklings imprint on the relational concept of same or different. *Science (80-.).*, 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.

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.

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. pages 1–16.

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 (80-.).*, 171(3972):701–703.

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.

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