A Benchmark for Compositional Visual Reasoning

Aimen Zerroug\textsuperscript{1,2,3}, Mohit Vaishnav\textsuperscript{1,2,3}, Julien Colin\textsuperscript{2}, Sebastian Musslick\textsuperscript{2}, Thomas Serre\textsuperscript{1,2,3}

\textsuperscript{1} Artificial and Natural Intelligence Toulouse Institute, Université de Toulouse, France
\textsuperscript{2} Carney Institute for Brain Science, Dpt. of Cognitive Linguistic & Psychological Sciences
Brown University, Providence, RI 02912
\textsuperscript{3} Centre de Recherche Cerveau et Cognition, CNRS, Université de Toulouse, France
aimen_zerroug@brown.edu

Abstract

A fundamental component of human vision is our ability to parse complex visual scenes and judge the relations between their constituent objects. AI benchmarks for visual reasoning have driven rapid progress in recent years with state-of-the-art systems now reaching human accuracy on some of these benchmarks. Yet, a major gap remains in terms of the sample efficiency with which humans and AI systems learn new visual reasoning tasks. Humans’ remarkable efficiency at learning has been at least partially attributed to their ability to harness compositionality – such that they can efficiently take advantage of previously gained knowledge when learning new tasks. Here, we introduce a novel visual reasoning benchmark, Compositional Visual Relations (CVR), to drive progress towards the development of more data-efficient learning algorithms. We take inspiration from fluidic intelligence and non-verbal reasoning tests and describe a novel method for creating compositions of abstract rules and associated image datasets at scale. Our proposed benchmark includes measures of sample efficiency, generalization and transfer across task rules, as well as the ability to leverage compositionality. We systematically evaluate modern neural architectures and find that convolutional architectures surpass transformer-based architectures across all performance measures in most data regimes. However, all computational models are much less data efficient than humans even after learning informative visual representations using self-supervision. Overall, we hope our challenge will spur interest in developing neural architectures that can learn to harness compositionality toward more efficient learning.

1 Introduction

Visual reasoning is a rather complex ability considering the high dimensionality of the sensory input and the level of abstraction it requires. It highlights human’s capacity to manipulate concepts and relations as symbols extracted from the visual input. The efficiency with which humans learn new visual concepts and relations, as exemplified by fluidic intelligence and non-verbal reasoning tests is equally fascinating. In the pursuit of human-level artificial intelligence, a growing body of research is attempting to emulate this skill in machines, and deep neural networks are at the forefront of the field. Deep learning approaches are prime candidates as models of human intelligence due to their success in learning from data while using simple design principles. However, these architectures are still

lacking due to the enormous amounts of data required for training, the inability to generalize to unfamiliar situations [13] and the lack of robustness [14]. Their ability to perform well in large-data regimes has skewed research in the field towards scaling up datasets and architectures with little consideration for the sample efficiency of these systems.

Only a few benchmarks address these aspects of human intelligence. Among them, ARC [9] provides diverse visual reasoning problems. However, the extreme scarcity of training samples, only 3 samples per task, renders the benchmark difficult for all methods, especially neural networks. Other benchmarks have guided progress in the field – helping spur the development of several neural network-based models [3, 43, 12]. Some focus on evaluating the task’s perceptual requirements [12], such as detecting features, recognizing objects, perceptual grouping and spatial reasoning. In contrast, others evaluate the logical reasoning requirements [3, 43], such as symbolic reasoning, making analogies and causal reasoning. However, they lack either the variety of abstract relations incorporated in the scene or the semantic and structural variety of scenes over which they instantiate these abstract relations.

Creating novel visual reasoning tasks can be challenging. In this benchmark, we standardize a process for creating tasks compositionally based on an elementary set of relations and abstractions. This process allows us to exploit a wide range of visual relations as well as abstract rules, thus, making it possible to evaluate both the perceptual and logical requirements of visual reasoning. Interestingly, the compositional nature of the tasks gives an opportunity to investigate the learning strategies wielded by existing methods. Among these methods, we focus on state-of-the-art abstract visual reasoning models and standard vision models. These models have been shown to reach high performance on several visual reasoning tasks in previous works [40, 38], but they always require large amounts of data. This paper’s subject of interest is quantifying these models’ sample efficiency.

**Contributions**

Our contributions can be summarized as follows:

- A novel visual reasoning benchmark **Compositional Visual Relations** (CVR) with 103 unique task rules over distinct scene structures.
- A novel method for generating visual reasoning problems with compositionality prior.
- A systematic analysis of the sample efficiency of baseline visual reasoning architectures.
- An empirical study of models’ capacity at harnessing compositionality to solve complex problems.

Our large-scale experiments capture a multitude of setups, including training on joint and individual dataset tasks, pre-training with self-supervision on dataset images to contrast learning of visual representations vs. abstract visual reasoning rules, and training over a range of data regimes and tests for transfer learning between dataset tasks. We present an in-depth analysis of task difficulty, which provides insights into the strengths and weaknesses of current models. Overall, we find that the best baselines trained in the most favorable conditions fall short of human sample efficiency for learning.
those same tasks. While models appear to be capable of transferring knowledge across tasks, we show that they do not leverage compositionality to efficiently learn task components. We hope to inspire research on more efficient visual reasoning models by releasing our dataset. The code for generating the full dataset and training models is available [here](#).

## 2 Compositional Visual Relations Dataset

CVR is a synthetic visual reasoning dataset that builds on prior AI benchmarks [12, 9] and on a body of cognitive science literature [37] on visual reasoning. In the following, we will describe the generation process of the dataset problems.

### Odd-One-Out

The odd one out task has been employed in prior work to test visual reasoning [27]. A sample problem consists of 4 images generated such that one of them is an outlier according to a certain rule. The goal of the task is to select the outlier. The learner is expected to test several hypotheses in order to detect the outlier. This process requires them to infer the hidden scene structure and relationships between the objects.

### Scene generation

Each image contains one **scene** composed of multiple **objects** as shown in Figure 2. An object is defined as a closed contours with a set of **attributes**: shape, position, size, color, rotation and flip. Other attributes describe the scene or low-level relations between objects. **Count** corresponds to the number of objects, groups of objects or relations. **Insideness** indicates that an object contains another object within its contour. **Contact** indicates that two object contours are touching. These 9 attributes are the basis for the 9 **elementary relations**. For example, a "size" relation is a constraint on the sizes of certain objects in the scene. Relations are expressed with natural language or

### Algorithm 1: Problem Generation Program

Generates problem samples of the shape-size task in Figure 2.

\[
\begin{align*}
& n \leftarrow 4 \quad \text{// Number of objects} \\
& \text{for } i \leftarrow 1 \text{ to } 4 \text{ do} \\
& \quad s \leftarrow \text{sample_size}() \\
& \quad s' \leftarrow s \times \text{rand}([2/3, 1/4]) \\
& \quad \text{if } i = 4 \text{ then} \\
& \quad \quad /\!\!/ \text{ Odd-One-Out} \\
& \quad \quad [s_1]^{1-n} \leftarrow [s, s', s, s'] \\
& \quad \quad \text{else} \\
& \quad \quad [s_1]^{1-n} \leftarrow [s, s, s', s'] \\
& \quad \text{end} \\
& \quad [o, o'] \leftarrow \text{sample_shapes}(n = 2) \\
& \quad [a_1]^{1-n} \leftarrow [o, o', o'] \\
& \quad [p_1]^{1-n} \leftarrow \text{sample_position}([s_1]^{1-n}) \\
& \quad [c_1]^{1-n} \leftarrow \text{sample_color}(n = 1) \\
& \text{end} \\
& \text{scene}_{1-4} = [o, p, s, c]^{1-n}_{1-4} \\
& \text{image}_{1-4} = \text{render}(\text{scene})_{1-4}
\end{align*}
\]
The larger object always has the same color

Position-Rotation
2 rotations of the same group of objects

Flip-Color
Flips of the same object have same color

Counting-Contact
The number of objects in contact in each group is constant

Size-Inside
The bigger object contains an object

Figure 3: Examples of task rules that are composed from a pair of relations. More examples are provided in the SI.

Rules and problem creation

The generation process described above can be used to instantiate different tasks; binary classification, few-shot binary classification, or a raven’s progressive matrix.

In this paper, we choose to apply this process to create odd-one-out problems. First, the task designer selects target relations and incorporates them into a new scene structure. In Fig. 2, the target relations are size and shape similarity; they are added to a scene with 4 objects. Then, a reference rule and an odd rule are chosen such that they combine target relations in different ways. The reference and odd rules in the example vary only in the size or shape attributes. A valid odd-one-out rule contradicts the reference rule such that any strategy used for solving the task must involve reasoning over the target relations exclusively. Given a scene structure, a reference and an odd-one-out rule, the generation process has a set of free parameters that control the generation process for new samples. The problem’s difficulty level can be varied by randomizing or fixing these parameters. In the shape-size task, the range of color values and the variation of objects across the 4 images are examples of free parameters. A higher number of random parameters results in a higher difficulty. We create generalization test sets by changing the sets of fixed or random parameters. For more details on the generalization test sets we refer the reader to the SI.

Dataset details

CVR incorporates 103 unique reference rules, including 9 rules built on the 9 elementary visual relations and 94 additional rules built on compositions of the rules. These compositions span all pairs of elementary rules and include up to 4 relations. While certain rules are composed of the same elementary relations, they remain unique in their scene structure or associations with other relations. The procedural generation of problem samples helps us create an arbitrary number of samples. We provide 10,000 training problem samples, 500 validation samples and 1,000 test samples for each task. We also provide a generalization test set of 1000 samples.

The compositionality prior is a design constraint for rules which ensures that solving the task requires reasoning over specific relations between objects in the visual scene. In the size-shape task the task shown in figure 2 the outlier can be differentiated from the other images by reasoning purely on
Table 1: Performance comparison: For each model, we report the accuracy and number of tasks with accuracy above 80%. SES is the Sample Efficiency Score; it favors models with high performance in low data regimes and consistent accuracy across regimes. SES and AUC are computed over the 20-1000 data regimes. OOD Generalization results are provided in the SI.

<table>
<thead>
<tr>
<th>N train samples</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>500</th>
<th>1000</th>
<th>SES</th>
<th>AUC</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50</td>
<td>28.0</td>
<td>31.1</td>
<td>32.5</td>
<td>34.0</td>
<td>38.7</td>
<td>44.8</td>
<td>33.7</td>
<td>34.9</td>
<td>-</td>
</tr>
<tr>
<td>ViT-small</td>
<td>28.6</td>
<td>30.1</td>
<td>30.9</td>
<td>31.9</td>
<td>33.8</td>
<td>35.1</td>
<td>31.3</td>
<td>31.7</td>
<td>-</td>
</tr>
<tr>
<td>SCL</td>
<td>26.9</td>
<td>30.0</td>
<td>30.3</td>
<td>30.0</td>
<td>31.4</td>
<td>33.4</td>
<td>31.3</td>
<td>31.7</td>
<td>-</td>
</tr>
<tr>
<td>WReN</td>
<td>30.0</td>
<td>32.0</td>
<td>32.9</td>
<td>34.1</td>
<td>36.3</td>
<td>39.0</td>
<td>33.4</td>
<td>34.1</td>
<td>-</td>
</tr>
<tr>
<td>SCL-ResNet 18</td>
<td>31.4</td>
<td>37.7</td>
<td>39.6</td>
<td>42.7</td>
<td>48.3</td>
<td>38.4</td>
<td>39.5</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

| rank-init       | 28.0| 31.1| 32.5| 34.0| 38.7| 44.8 | 33.7| 34.9| -    |
| ResNet-50       | 27.5| 28.2| 29.9| 33.9| 52.1| 59.2 | 59.2| 59.2| -    |
| ViT-small       | 27.3| 27.8| 28.1| 31.4| 38.4| 57.0 | 57.0| 57.0| -    |
| SCL             | 25.8| 25.8| 28.3| 34.1| 43.2| 54.8 | 54.8| 54.8| -    |
| WReN            | 26.8| 27.6| 28.5| 30.1| 36.4| 40.8 | 40.8| 40.8| -    |
| SCL-ResNet 18   | 26.4| 28.4| 31.6| 40.7| 64.0| 47.1 | 47.1| 47.1| -    |

| ind  | 40.5| 47.3| 52.9| 56.8| 61.9| 67.7 | 52.4| 54.5| -    |
| ResNet-50       | 46.7| 51.6| 54.8| 57.5| 62.0| 65.5 | 54.9| 56.4| -    |
| ViT-small       | 27.9| 28.2| 28.6| 30.0| 35.6| 47.2 | 31.7| 32.9| -    |
| SCL             | 26.8| 32.0| 40.8| 44.1| 45.4| 49.7 | 32.0| 34.1| -    |
| WReN            | 31.1| 37.4| 43.9| 56.0| 68.9| 78.8 | 48.9| 52.7| -    |

| joint | ResNet-50       | 44.3| 50.3| 55.3| 59.5| 68.9| 79.2 | 57.0| 59.6| 93.1 |
|       | ViT-small       | 39.3| 39.5| 40.8| 44.1| 53.3| 60.7 | 44.7| 46.3| 81.6 |
|       | SCL             | 32.0| 35.1| 39.0| 43.8| 57.7| 69.5 | 43.4| 46.2| -    |
|       | WReN            | 27.9| 28.2| 28.6| 30.0| 35.6| 47.2 | 31.7| 32.9| -    |
|       | SCL-ResNet 18   | 28.7| 32.0| 40.8| 44.1| 45.4| 49.7 | 32.0| 34.1| -    |
|       | ViT-base        | 31.1| 37.4| 43.9| 56.0| 68.9| 78.8 | 48.9| 52.7| -    |

| SSL  | ResNet-50       | 27.5| 28.2| 29.9| 33.9| 52.1| 59.2 | 59.2| 59.2| -    |
| ind  | ViT-small       | 27.3| 27.8| 28.1| 31.4| 38.4| 57.0 | 57.0| 57.0| -    |
| SCL  | 25.8| 25.8| 28.3| 34.1| 43.2| 54.8 | 54.8| 54.8| -    |
| WReN | 26.8| 27.6| 28.5| 30.1| 36.4| 40.8 | 40.8| 40.8| -    |
| SCL-ResNet 18 | 26.4| 28.4| 31.6| 40.7| 64.0| 47.1 | 47.1| 47.1| -    |

| joint | ResNet-50       | 44.3| 50.3| 55.3| 59.5| 68.9| 79.2 | 57.0| 59.6| 93.1 |
|       | ViT-small       | 39.3| 39.5| 40.8| 44.1| 53.3| 60.7 | 44.7| 46.3| 81.6 |
|       | SCL             | 32.0| 35.1| 39.0| 43.8| 57.7| 69.5 | 43.4| 46.2| -    |
|       | WReN            | 27.9| 28.2| 28.6| 30.0| 35.6| 47.2 | 31.7| 32.9| -    |
|       | SCL-ResNet 18   | 28.7| 32.0| 40.8| 44.1| 45.4| 49.7 | 32.0| 34.1| -    |
|       | ViT-base        | 31.1| 37.4| 43.9| 56.0| 68.9| 78.8 | 48.9| 52.7| -    |

3 Experimental setting

Baseline models In our experiments, we select two standard vision models commonly used in computer vision. We evaluate ResNet [15], a convolutional architecture used as a baseline in several benchmarks [3,43,38] and also used as a backbone in standard VQA models. As a transformer-based architecture, we choose ViT [11]. It is used in various vision tasks such as image classification,
object recognition, image captioning and recently in visual reasoning on SVRT [28]. To evaluate the architectures on fairgrounds, we choose ResNet-50 and ViT-small, which have an equal number of parameters. Additionally, we evaluate two baseline visual reasoning models designed for solving RPMs: SCL [40] which boasts state-of-the-art accuracy on RAVEN and PGM, and WReN [3] which is based on a relational reasoning model [33]. Finally, we present a model that combines ResNet’s perception with SCL’s reasoning skills.

<table>
<thead>
<tr>
<th>N training samples</th>
<th>20</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50</td>
<td>28.0</td>
<td>57.9</td>
</tr>
<tr>
<td>ViT-small</td>
<td>29.3</td>
<td>32.7</td>
</tr>
<tr>
<td>SCL</td>
<td>26.4</td>
<td>44.9</td>
</tr>
<tr>
<td>WReN</td>
<td>27.5</td>
<td>42.4</td>
</tr>
<tr>
<td>SCL-ResNet 18</td>
<td>26.8</td>
<td>64.1</td>
</tr>
<tr>
<td>ResNet-50 SSL</td>
<td>45.7</td>
<td>7</td>
</tr>
<tr>
<td>ViT-small SSL</td>
<td>38.7</td>
<td>60.3</td>
</tr>
<tr>
<td>Humans</td>
<td>78.7</td>
<td>26</td>
</tr>
</tbody>
</table>

Table 2: Human Baseline: performance of models on joint training experiments is compared to the human baseline. The analysis is restricted to the 45 tasks used for evaluating humans. ResNet 50 approaches human-level performance only after SSL pre-training and finetuning on all task rules with 1000 samples per rule. Which is 50 times higher than the number of samples needed by humans.

Joint vs. Individual rule learning Models are trained on one task or jointly on all tasks. In the context of CVR, the random generation of scenes chooses an image as an odd-one-out with respect to the reference rule. Still, it might also cause another image to be an outlier with respect to the irrelevant rules. To illustrate this problem, let’s take the elementary size rule as an example. In this rule, each image contains one object. Due to the random sampling of object attributes, it is possible for one image to be considered an outlier with respect to the color rule (The attributes in the 4 images are i-small/green, ii-large/green, iii-small/green, iv-small/blue). Without specifying that the rule to solve is a size rule, the model could incorrectly choose the fourth image because it is an outlier with respect to the color rule. Thus, models trained on several tasks could easily confound rules. To avoid this problem, they are provided with a rule embedding vector. Given the rule token, models can learn several strategies and use the correct one for each problem samples. We also compare the two settings as they allow for testing the model’s capacity and efficiency at learning several strategies and routines to solve different rules. All hyperparameter choices and training details are provided in the SI.

Self-Supervised pre-training Unlike humans who spend a lifetime analyzing visual information, randomly initialized neural networks have no visual experience. To provide a fair comparison between humans and neural networks, we pre-train baseline models on a subset of the training data. Self-Supervised Learning (SSL) has seen a rise in popularity due to its usefulness in pre-training models on unlabelled data. In our setting, by using SSL, we aim to dissociate feature learning from abstract visual reasoning in standard vision models. We pre-trained ViT-small and ResNet-50 on 1 million images from the dataset following MoCo-v3 [8]. In addition to SSL pre-trained models, we also finetune models pre-trained on object recognition and image annotation. As a natural task that humans perform regularly, image annotation requires visual reasoning capabilities and would provide a fair comparison with humans. We select ResNet-50 and ViT-small pre-trained on ImageNet [10]. We also pick CLIP [31] visual encoders ResNet-50 and ViT-Base that are trained jointly with a language model on image annotation.

Human Baseline As found in [12], having 21 participants solve the 9 task rules based on elementary relations and 36 randomly sampled complex task rules is sufficient to yield a reliable human baseline. We used 20 problem samples for each rule which corresponds to the lowest number of samples used for training baseline models. Each participant performed 6 different rules. More details about the behavioral experiment are provided in the SI.

4 Results

Sample Efficiency Baseline models are trained in six data regimes ranging from 20 to 1000 training samples. All sample efficiency results are summarized in Table 1. The random guess accuracy level
in the dataset is 25%. We observe that most randomly initialized models are slightly above the random guess accuracy after training in low data regimes. They achieve an increase in performance only when provided with more than 500 training samples. SCL-ResNet-18, followed by ResNet-50, performs the best in high data regimes, while SCL and ViT have the lowest performance in high data regimes. This result was expected since transformer architectures generally learn better in high data regimes (millions of data points) and are consistent with prior work [38] which finds that ViTs do not learn several SVRT tasks even when trained on 100k samples. Although SCL’s performance is near chance, when augmented with a strong vision backbone, ResNet-18, it achieves the best performance. This jump in performance is indicative of the two architectures’ complementary roles in visual reasoning. Results in Table 1 and Fig. 4 show a clear positive effect for pre-training on all models. SSL pre-trained achieve the highest performance compared to object recognition and image annotation pre-trained models. We observe that ViT benefits from a larger architecture coupled with pre-training on a large image annotation dataset. This highlights transformers’ reliance on large model sizes and datasets.

In order to quantify sample efficiency systematically for all models, we compute the area under the curve (AUC), which corresponds to the unweighted average performance across data regimes. We also introduce Sample Efficiency Score (SES) as an empirical evaluation metric for our experimental setting. It consists of a weighted average of accuracy where the weights are reversely proportional to number of samples: 

$$SES = \frac{1}{n} \sum_{i=1}^{n} \log(n) \cdot w_n$$

where

$$w_n = \frac{1}{1 + \log(n)}$$

and $n$ is the number of samples. This score favors models that learn with the fewest samples while considering consistency in the overall performance. We observe that SCL-ResNet 18 scores the highest in the individual and joint training settings. In the SSL finetuning condition, ViT and ResNet-50 have a similar SES when trained on individual tasks, but ResNet-50 performs better in the joint training setting. These results hint at the efficiency of convolutional architectures in visual reasoning tasks. The best performance on CVR is 93.7%, achieved by ResNet-50 in the joint-rule learning setting and the 10k data regime. This high performance levels in the 10,000 data regime demonstrate the models’ capacity to learn the majority of rules in the dataset and strongly highlights that failure in lower data regimes is explained by their sample inefficiency.

Finally, we compare model performance to the human baseline. We observe in Table 2 that humans far exceed the accuracy of all models with only 20 samples. This result is supported by previous work on the SVRT dataset [12] where participants solved similar tasks with less than 20 samples. These results highlight the gap between humans and machines in sample efficiency and encourage the development of more efficient architectures.

**Compositionality** Transferring knowledge and skills across tasks is a crucial feature of intelligent systems. With our experimental setup, this can be characterized in several ways. A compositional model should reuse acquired skills to learn efficiently. Thus, when it is trained on all rules jointly, it should be more sample efficient because the rules in the dataset share elementary components. In Table 1 and Figure 4, we observe that ResNet-50 achieves higher performance on joint training compared to individual rule training while ViT has an opposite effect. The trend is consistent across data regimes and other settings. These results highlight convolutional architectures’ learning efficiency compared to transformer architectures.
We investigate compositionality further by asking whether learning elementary rules provides a better initialization for learning their compositions. For example, a model that can judge object positions and sizes should not require many training samples to judge objects’ relative positions identified by their sizes. We pick a set of rules with at least two different components, train models to reach the maximum accuracy possible on component relations, then finetune the models on the compositions. We call this experimental condition the curriculum condition since the condition is akin to incrementally teaching routines to a model. Model performance in the curriculum condition is compared to performance when trained from scratch. The results highlighted in Fig. 5a show positive effects for most models but more significantly for convolution-based architectures. These results indicate that the baselines use skills acquired during pre-training to learn the composition rule to varying degrees. We refer the readers to the SI for additional analyses and quantitative results.

Finally, we evaluate transfer learning from composition rules to elementary rules. We name this condition the reverse curriculum condition. The working hypothesis is that models that rely on compositionality will be able to solve elementary relations without finetuning if they learn the composition. We compare accuracy on the composition rule to the zero-shot accuracy on the respective elementary rules in Fig. 5b. We observe that all models perform worse on the elementary relations. These results might indicate that although the baselines could transfer skills from elementary rules to their compositions, they do not necessarily use an efficient strategy that decomposes tasks into their elementary components. Additional analyses are presented in the SI.

**Task difficulty** We analyze the performance of all models in the standard setting: joint training on all rules from random initialization. Fig. 6 shows the average performance of each model on each elementary rule composition rule. Since the dataset contains several compositions of each pair of elementary rules, the accuracy showed in each square is averaged over rules that share the same pair of elementary rules. Certain rules are solvable by all models, such as the position, size, color, and count elementary rules. Additionally, certain rules pose a challenge for all models, these rules are compositions of count, flip, rotation or shape. Models that rely on a convolutional backbone were able to solve most spatial rules; position, size, inside and contact. However, they fail on rules that incorporate shapes and their transformations; shape, rotation, flip. Composition rules built with the
Figure 6: Rule differences The performance at 1000 samples is shown for each model in the joint training setting. Performance on elementary rules is shown on the top row. Performance on compositions is indexed by the annotations on the diagonal. We observe that all models fail at most compositions based on "color".

Count relation proved to be a challenge for most models. We believe that models are capable of solving several tasks, such as the counting elementary rule, by relying on shortcuts; this could be a summation of all pixels in the image, for example. These shortcuts prevent models from learning abstract rules and hinder generalization. In line with the previous results, SCL-ResNet-18 seems to solve more elementary rules and compositions than the other 3 models.

5 Related Work

Visual reasoning benchmarks Visual reasoning has been a subject of AI research for decades, and several benchmarks address many relevant tasks. This includes language-guided reasoning benchmarks such as CLEVR [18], extended in its visual composition by recent work [23], physics-based reasoning and reasoning over time dynamics [42,4]. More relevant to our work are abstract visual reasoning benchmarks. Raven’s Progressive Matrices (RPMs) are one example introduced in 1938 [6] to test human fluidic intelligence. Procedural generation techniques for RPMs [59] enabled the creation of the PGM dataset and RAVEN [4,53]. They also inspired Bongard-Logo [29], a concept learning and reasoning benchmark based on Bongard’s 100 visual reasoning problems [4]. Another reasoning dataset, SVRT [12], focuses on evaluating similarity-based judgment and spatial reasoning. Besides these synthetic datasets, real-world datasets were developed with similar task structures to Bongard-Logo and RPM [55,17]. In this work, we take inspiration from SVRT and develop a more extensive set of rules with careful considerations for the choice of rules and using a novel rule generation method. Finally, Abstract Reasoning Corpus [9] is a general intelligence test introduced with a new methodology for evaluating intelligence and generalization. The numerous problems presented in this benchmark are constructed with a variety of human priors. The unique nature of the task, requiring solvers to generate the answer, and the limited amount of training data render the benchmark difficult for neural network-based methods. We follow a similar approach in our dataset by creating several unique problem templates. However, we restrict the number of samples to a reasonable range to evaluate the sample efficiency of candidate models.

Compositionality Compositionality is a highly studied topic in AI research. Although there is agreement over the high-level definition of compositionality; the ability to represent new abstractions based on their constituents and their contexts, there is little consensus over methods for characterizing compositional generalization in neural networks. Several tests for compositionality have been proposed in language [26], mathematics [33], logical reasoning and navigation [5,21,32,41] and visual reasoning [18,36,1]. Recent work [16] attempts to identify components of compositionality and proposes a test suit that unifies them. These tests evaluate the model’s capacity to manipulate concepts during inference; systematicity tests the novel combination of features, akin to CLEVR’s CoGenT [18] and C-VQA [1] where novel combinations of shapes and colors introduced in the test set, and localism tests the model’s ability to account for context similarly to samples from Winoground [36]. Our work explores compositional generalization from a new perspective; CVR evaluates the model’s compositionality while learning novel concepts. A compositional model reuses
previously learned concepts to accelerate learning and decomposes complex tasks into elementary components. These aspects of compositionality are tested under settings that employ curricula. Furthermore, we evaluate compositionality over the reasoning operations necessary to solve a given problem. Finally, generating a synthetic dataset allows for evaluating reasoning at high levels of abstraction; groups of objects and scene configurations as exemplified by tasks in Fig 3.

Neuroscience/Psychology Several theories attempt to propose an understanding of the mechanisms behind visual reasoning, such as gestalt psychology which outlines principles hypothesized to be used by the visual system as an initial set of abstractions. Another theory describes visual reasoning as a sequence of elemental operations called visual routines [37] orchestrated by higher-level cognitive processes. These elemental operations are hypothesized to form the basis for spatial reasoning, same-different judgment, perceptual grouping, contour tracing and many other visual skills [7]. Evaluating these skills in standard vision models is a recurring subject in machine learning and neuroscience research [19, 24, 30]. To provide a comprehensive evaluation of visual reasoning, it is important to include task sets that require various visual skills within humans’ capabilities.

6 Discussion and Future Work

In this work, we have proposed a novel benchmark that focuses on two important aspects of human intelligence – compositionality and sample efficiency. Inspired by visual cognition theories [37], the proposed challenge addresses the limitations of existing benchmarks in the following ways: (1) it extends previous benchmarks by providing a variety of visual reasoning tasks that vary in relations and scene structures, (2) all tasks in the benchmark were designed with a compositionality prior, which allows for an in-depth analysis of each model’s strengths and weaknesses, and (3) it provides a quantitative measure of sample efficiency.

Using this benchmark, we performed an analysis of the sample efficiency of existing machine learning models and their ability to harness compositionality. Our results suggest that even the best pre-trained neural architectures require magnitudes more training samples than humans to reach the same level of accuracy, which is consistent with prior work on sample efficiency [22]. Our evaluation further revealed that current neural architectures fail to learn several tasks even when provided an abundance of samples and extensive prior visual experience. These results highlight the importance of developing more data-efficient and vision-oriented neural architectures toward achieving human-level artificial intelligence. In addition, we evaluated models’ generalization ability across rules – from elementary rules to compositions and vice versa. We find that convolutional architectures benefit from learning all visual reasoning tasks jointly and transferring skills learned during training on elementary rules. However, they also failed to generalize systematically from compositions to their individual rules. These results indicate that convolutional architectures are capable of transferring skills across tasks but do not learn by decomposing a visual task into its elementary components.

While our work addresses important questions on sample efficiency and compositionality, we note a few possible limitations of our proposed benchmark. CVR is quite extensive in terms of the visual relations it contains, but it can always be further improved in its use of elementary visual relations. For example, the shapes could be parametrically generated based on specific geometric features. Hopefully, CVR can be expanded in future work to test more routines by including additional relations borrowed from other, more narrow challenges, including occlusion [19], line tracing [25], and physics-based relations. The rules in the current benchmark are limited to 2 or 3 levels of abstraction to evaluate relations systematically. Similarly, our evaluation methods for sample efficiency and compositionality could be further improved and adapted to different settings. For example, the sample efficiency score is an empirical metric used only for evaluating our benchmark. It requires training all models on all data regimes for the score to be consistent. Although our work is not unique in addressing sample efficiency, its aim is promoting more sample efficient and general models. We hope that the release of our benchmark will encourage researchers in the field to test their own model’s sample efficiency and compositionality.
7 Acknowledgments

This work is funded by NSF (IIS-1912280) and ONR (N00014-19-1-2029) to TS. Additional support was provided by the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANR-19-P13A-0004). Computing resources used supported by the Center for Computation and Visualization (NIH S10OD025181) at Brown and CALMIP supercomputing center (Grant 2016-p20019, 2016-p22041) at Federal University of Toulouse Midi-Pyrénées.

References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Limitations are discussed in the paper and expanded in the supplementary materials.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] The social impact of the paper discussed in the supplementary material
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All training details are specified in the supplementary material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We perform hyperparameter search in preliminary experiments and perform the main experiments with one seed and one set of hyperparameters because they are computationally expensive.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes]