

A simple neural network module for relational reasoning

Adam Santoro*, David Raposo*, David G.T. Barrett, Mateusz Malinowski,
Razvan Pascanu, Peter Battaglia, Timothy Lillicrap

adamsantoro@, draposo@, barrettdavid@, mateuszm@,
razp@, peterbattaglia@, countzero@google.com

DeepMind
London, United Kingdom

Abstract

Relational reasoning is a central component of generally intelligent behavior, but has proven difficult for neural networks to learn. In this paper we describe how to use Relation Networks (RNs) as a simple plug-and-play module to solve problems that fundamentally hinge on relational reasoning. We tested RN-augmented networks on three tasks: visual question answering using a challenging dataset called CLEVR, on which we achieve state-of-the-art, super-human performance; text-based question answering using the bAbI suite of tasks; and complex reasoning about dynamic physical systems. Then, using a curated dataset called Sort-of-CLEVR we show that powerful convolutional networks do not have a general capacity to solve relational questions, but can gain this capacity when augmented with RNs. Our work shows how a deep learning architecture equipped with an RN module can implicitly discover and learn to reason about entities and their relations.

1 Introduction

The ability to reason about the relations between entities and their properties is central to generally intelligent behavior (Figure 1) [18, 15]. Consider a child proposing a race between the two trees in the park that are furthest apart: the pairwise distances between every tree in the park must be inferred and compared to know where to run. Or, consider a reader piecing together evidence to predict the culprit in a murder-mystery novel: each clue must be considered in its broader context to build a plausible narrative and solve the mystery.

Symbolic approaches to artificial intelligence are inherently relational [32, 11]. Practitioners define the relations between symbols using the language of logic and mathematics, and then reason about these relations using a multitude of powerful methods, including deduction, arithmetic, and algebra. But symbolic approaches suffer from the symbol grounding problem and are not robust to small task and input variations [11]. Other approaches, such as those based on statistical learning, build representations from raw data and often generalize across diverse and noisy conditions [25]. However, a number of these approaches, such as deep learning, often struggle in data-poor problems where the underlying structure is characterized by sparse but complex relations [7, 23]. Our results corroborate these claims, and further demonstrate that seemingly simple relational inferences are remarkably

*Equal contribution.

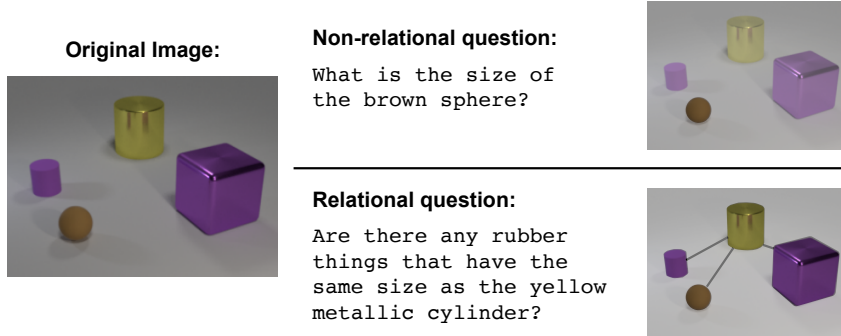


Figure 1: **An illustrative example from the CLEVR dataset of relational reasoning.** An image containing four objects is shown alongside non-relational and relational questions. The relational question requires explicit reasoning about the relations between the four objects in the image, whereas the non-relational question requires reasoning about the attributes of a particular object.

difficult for powerful neural network architectures such as convolutional neural networks (CNNs) and multi-layer perceptrons (MLPs).

Here, we explore “Relation Networks” (RN) as a general solution to relational reasoning in neural networks. RNs are architectures whose computations focus explicitly on relational reasoning [35]. Although several other models supporting relation-centric computation have been proposed, such as Graph Neural Networks, Gated Graph Sequence Neural Networks, and Interaction Networks, [37, 26, 2], RNs are simple, plug-and-play, and are exclusively focused on flexible relational reasoning. Moreover, through joint training RNs can influence and shape upstream representations in CNNs and LSTMs to produce implicit object-like representations that it can exploit for relational reasoning. We applied an RN-augmented architecture to CLEVR [15], a recent visual question answering (QA) dataset on which state-of-the-art approaches have struggled due to the demand for rich relational reasoning. Our networks vastly outperformed the best generally-applicable visual QA architectures, and achieve state-of-the-art, super-human performance. RNs also solve CLEVR from state descriptions, highlighting their versatility in regards to the form of their input. We also applied an RN-based architecture to the bAbI text-based QA suite [41] and solved 18/20 of the subtasks. Finally, we trained an RN to make challenging relational inferences about complex physical systems and motion capture data. The success of RNs across this set of substantially dissimilar task domains is testament to the general utility of RNs for solving problems that require relation reasoning.

2 Relation Networks

An RN is a neural network module with a structure primed for relational reasoning. The design philosophy behind RNs is to **constrain the functional form of a neural network so that it captures the core common properties of relational reasoning**. In other words, the capacity to compute relations is baked into the RN architecture without needing to be learned, just as the capacity to reason about spatial, translation invariant properties is built-in to CNNs, and the capacity to reason about sequential dependencies is built into recurrent neural networks.

In its simplest form the RN is a composite function:

$$\text{RN}(O) = f_{\phi} \left(\sum_{i,j} g_{\theta}(o_i, o_j) \right), \quad (1)$$

where the input is a set of “objects” $O = \{o_1, o_2, \dots, o_n\}$, $o_i \in \mathbb{R}^m$ is the i^{th} object, and f_{ϕ} and g_{θ} are functions with parameters ϕ and θ , respectively. For our purposes, f_{ϕ} and g_{θ} are MLPs, and the

parameters are learnable synaptic weights, making RNs end-to-end differentiable. We call the output of g_θ a “relation”; therefore, **the role of g_θ is to infer the ways in which two objects are related**, or if they are even related at all.

RNs have three notable strengths: they learn to infer relations, they are data efficient, and they operate on a set of objects – a particularly general and versatile input format – in a manner that is order invariant.

RNs learn to infer relations The functional form in Equation 1 dictates that an RN should consider the potential relations between *all* object pairs. This implies that an RN is not necessarily privy to which object relations actually exist, nor to the actual meaning of any particular relation. Thus, RNs must *learn to infer* the existence and implications of object relations.

In graph theory parlance, the input can be thought of as a complete and directed graph whose nodes are objects and whose edges denote the object pairs whose relations should be considered. Although we focus on this “all-to-all” version of the RN throughout this paper, this RN definition can be adjusted to consider only some object pairs. Similar to Interaction Networks [2], to which RNs are related, RNs can take as input a list of only those pairs that should be considered, if this information is available. This information could be explicit in the input data, or could perhaps be extracted by some upstream mechanism.

RNs are data efficient RNs use a single function g_θ to compute each relation. This can be thought of as a single function operating on a batch of object pairs, where each member of the batch is a particular object-object pair from the same object set. This mode of operation encourages greater generalization for computing relations, since g_θ is encouraged not to over-fit to the features of any particular object pair. Consider how an MLP would learn the same function. An MLP would receive *all* objects from the object set simultaneously as its input. It must then learn and embed n^2 (where n is the number of objects) identical functions within its weight parameters to account for all possible object pairings. This quickly becomes intractable as the number of objects grows. Therefore, the cost of learning a relation function n^2 times using a single feedforward pass per sample, as in an MLP, is replaced by the cost of n^2 feedforward passes per object set (i.e., for each possible object pair in the set) and learning a relation function just once, as in an RN.

RNs operate on a set of objects The summation in Equation 1 ensures that the RN is invariant to the order of objects in the input. This invariance ensures that the RN’s input respects the property that sets are order invariant, and it ensures that the output is order invariant. Ultimately, this invariance ensures that the RN’s output contains information that is generally representative of the relations that exist in the object set.

3 Tasks

We applied RN-augmented networks to a variety of tasks that hinge on relational reasoning. To demonstrate the versatility of these networks we chose tasks from a number of different domains, including visual QA, text-based QA, and dynamic physical systems.

3.1 CLEVR

In visual QA a model must learn to answer questions about an image (Figure 1). This is a challenging problem domain because it requires high-level scene understanding [1, 29]. Architectures must perform complex relational reasoning – spatial and otherwise – over the features in the visual inputs, language inputs, and their conjunction. However, the majority of visual QA datasets require reasoning in the absence of fully specified word vocabularies, and perhaps more perniciously, a vast and complicated knowledge of the world that is not available in the training data. They also contain ambiguities and exhibit strong linguistic biases that allow a model to learn answering strategies that exploit those biases, without reasoning about the visual input [1, 31, 36].

To control for these issues, and to distill the core challenges of visual QA, the CLEVR visual QA dataset was developed [15]. CLEVR contains images of 3D-rendered objects, such as spheres and cylinders (Figure 2). Each image is associated with a number of questions that fall into different categories. For example, **query attribute** questions may ask “*What is the color of the sphere?*”, while **compare attribute** questions may ask “*Is the cube the same material as the cylinder?*”.

For our purposes, an important feature of CLEVR is that many questions are explicitly relational in nature. Remarkably, powerful QA architectures [46] are unable to solve CLEVR, presumably because they cannot handle core relational aspects of the task. For example, as reported in the original paper a model comprised of ResNet-101 image embeddings with LSTM question processing and augmented with stacked attention modules vastly outperformed other models at an overall performance of 68.5% (compared to 52.3% for the next best, and 92.6% human performance) [15]. However, for **compare attribute** and **count** questions (i.e., questions heavily involving relations across objects), the model performed little better than the simplest baseline, which answered questions solely based on the probability of answers in the training set for a given question category (Q-type baseline).

We used two versions of the CLEVR dataset: (i) the pixel version, in which images were represented in standard 2D pixel form, and (ii) a state description version, in which images were explicitly represented by state description matrices containing factored object descriptions. Each row in the matrix contained the features of a single object – 3D coordinates (x, y, z); color (r, g, b); shape (cube, cylinder, etc.); material (rubber, metal, etc.); size (small, large, etc.). When we trained our models, we used *either* the pixel version or the state description version, depending on the experiment, but not both together.

3.2 Sort-of-CLEVR

To explore our hypothesis that the RN architecture is better suited to general relational reasoning as compared to more standard neural architectures, we constructed a dataset similar to CLEVR that we call “Sort-of-CLEVR”¹. This dataset separates relational and non-relational questions.

Sort-of-CLEVR consists of images of 2D colored shapes along with questions and answers about the images. Each image has a total of 6 objects, where each object is a randomly chosen shape (square or circle). We used 6 colors (red, blue, green, orange, yellow, gray) to unambiguously identify each object. Questions are hard-coded as fixed-length binary strings to reduce the difficulty involved with natural language question-word processing, and thereby remove any confounding difficulty with language parsing. For each image we generated 10 relational questions and 10 non-relational questions. Examples of relational questions are: “*What is the shape of the object that is farthest from the gray object?*”; and “*How many objects have the same shape as the green object?*”. Examples of non-relational questions are: “*What is the shape of the gray object?*”; and “*Is the blue object on the top or bottom of the scene?*”. The dataset is also visually simple, reducing complexities involved in image processing.

3.3 bAbI

bAbI is a pure text-based QA dataset [41]. There are 20 tasks, each corresponding to a particular type of reasoning, such as deduction, induction, or counting. Each question is associated with a set of supporting facts. For example, the facts “*Sandra picked up the football*” and “*Sandra went to the office*” support the question “*Where is the football?*” (answer: “*office*”). A model succeeds on a task if its performance surpasses 95%. Many memory-augmented neural networks have reported impressive results on bAbI. When training jointly on all tasks using 10K examples per task, Memory Networks pass 14/20, DNC 18/20, Sparse DNC 19/20, and EntNet 16/20 (the authors of EntNets report state-of-the-art at 20/20; however, unlike previously reported results this was not done with joint training on all tasks, where they instead achieve 16/20) [42, 9, 34, 13].

¹The “Sort-of-CLEVR” dataset will be made publicly available online.

3.4 Dynamic physical systems

We developed a dataset of simulated physical mass-spring systems using the MuJoCo physics engine [40]. Each scene contained 10 colored balls moving on a table-top surface. Some of the balls moved independently, free to collide with other balls and the barrier walls. Other randomly selected ball pairs were connected by invisible springs or a rigid constraint. These connections prevented the balls from moving independently, due to the force imposed through the connections. Input data consisted of state descriptions matrices, where each ball was represented as a row in a matrix with features representing the RGB color values of each object and their spatial coordinates (x, y) across 16 sequential time steps.

The introduction of random links between balls created an evolving physical system with a variable number “systems” of connected balls (where “systems” refers to connected graphs with balls as nodes and connections between balls as edges). We defined two separate tasks: 1) infer the existence or absence of connections between balls when only observing their color and coordinate positions across multiple sequential frames, and 2) count the number of systems on the table-top, again when only observing each ball’s color and coordinate position across multiple sequential frames.

Both of these tasks involve reasoning about the relative positions and velocities of the balls to infer whether they are moving independently, or whether their movement is somehow dependent on the movement of other balls through invisible connections. For example, if the distance between two balls remains similar across frames, then it can be inferred that there is a connection between them. The first task makes these inferences explicit, while the second task demands that this reasoning occur implicitly, which is much more difficult. For further information on all tasks, including videos of the dynamic systems, see the supplementary information.

4 Models

In their simplest form RNs operate on *objects*, and hence do not explicitly operate on images or natural language. 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. Although the RN expects object representations as input, the semantics of what an object *is* need not be specified. Our results below demonstrate that the learning process induces upstream processing, comprised of conventional neural network modules, to produce a set of useful “objects” from distributed representations.

Dealing with pixels We used a CNN to parse pixel inputs into a set of objects. The CNN took images of size 128×128 and convolved them through four convolutional layers to k feature maps of size $d \times d$, where k is the number of kernels in the final convolutional layer. We remained agnostic as to what particular image features should constitute an object. So, after convolving the image, each of the d^2 k -dimensional cells in the $d \times d$ feature maps was tagged with an arbitrary coordinate indicating its relative spatial position, and was treated as an object for the RN (see Figure 2). This means that an “object” could comprise the background, a particular physical object, a texture, conjunctions of physical objects, etc., which affords the model great flexibility in the learning process.

Conditioning RNs with question embeddings The existence and meaning of an object-object relation should be question dependent. For example, if a question asks about a large sphere, then the relations between small cubes are probably irrelevant. So, we modified the RN architecture such that g_θ could condition its processing on the question: $a = f_\phi(\sum_{i,j} g_\theta(o_i, o_j, q))$. To get the question embedding q , we used the final state of an LSTM that processed question words. Question words were assigned unique integers, which were then used to index a learnable lookup table that provided embeddings to the LSTM. At each time-step, the LSTM received a single word embedding as input, according to the syntax of the English-encoded question.

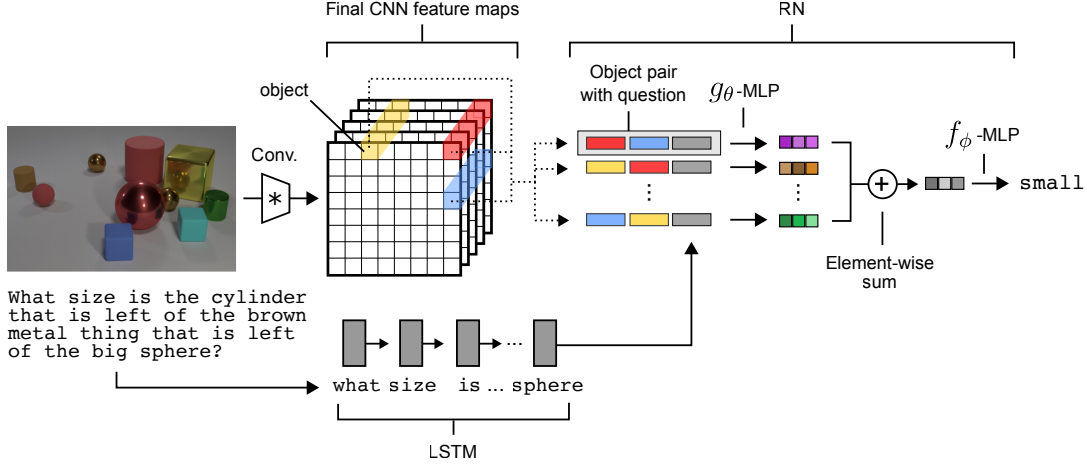


Figure 2: **Visual QA architecture.** Questions are processed with an LSTM to produce a question embedding, and images are processed with a CNN to produce a set of objects for the RN. Objects (three examples illustrated here in yellow, red, and blue) are constructed using feature-map vectors from the convolved image. The RN considers relations across all pairs of objects, conditioned on the question embedding, and integrates all these relations to answer the question.

Dealing with state descriptions We can provide state descriptions directly into the RN, since state descriptions are pre-factored object representations. Question processing can proceed as before: questions pass through an LSTM using a learnable lookup embedding for individual words, and the final state of the LSTM is concatenated to each object-pair.

Dealing with natural language For the bAbI suite of tasks the natural language inputs must be transformed into a set of objects. This is a distinctly different requirement from visual QA, where objects were defined as spatially distinct regions in convolved feature maps. So, we first identified up to 20 sentences in the support set that were immediately prior to the probe question. Then, we tagged these sentences with labels indicating their relative position in the support set, and processed each sentence word-by-word with an LSTM (with the same LSTM acting on each sentence independently). We note that this setup invokes minimal prior knowledge, in that we delineate objects as sentences, whereas previous bAbI models processed all word tokens from all support sentences sequentially. It’s unclear how much of an advantage this prior knowledge provides, since period punctuation also unambiguously delineates sentences for the token-by-token processing models. The final state of the sentence-processing-LSTM is considered to be an object. Similar to visual QA, a separate LSTM produced a question embedding, which was appened to each object pair as input to the RN. Our model was trained on the joint version of bAbI (all 20 tasks simultaneously), using the full dataset of 10K examples per task.

Model configuration details For the CLEVR-from-pixels task we used: 4 convolutional layers each with 24 kernels, ReLU non-linearities, and batch normalization; 128 unit LSTM for question processing; 32 unit word-lookup embeddings; four-layer MLP consisting of 256 units per layer with ReLU non-linearities for g_θ ; and a three-layer MLP consisting of 256, 256 (with 50% dropout), and 29 units with ReLU non-linearities for f_ϕ . The final layer was a linear layer that produced logits for a softmax over the answer vocabulary. The softmax output was optimized with a cross-entropy loss function using the Adam optimizer with a learning rate of $2.5e^{-4}$. We used size 64 mini-batches and distributed training with 10 workers synchronously updating a central parameter server. The configurations for the other tasks are similar, and can be found in the supplementary information.

We’d like to emphasize the simplicity of our overall model architecture compared to the visual QA architectures used on CLEVR thus far, which use ResNet or VGG embeddings, sometimes with

Model	Overall	Count	Exist	Compare Numbers	Query Attribute	Compare Attribute
Human	92.6	86.7	96.6	86.5	95.0	96.0
Q-type baseline	41.8	34.6	50.2	51.0	36.0	51.3
LSTM	46.8	41.7	61.1	69.8	36.8	51.8
CNN+LSTM	52.3	43.7	65.2	67.1	49.3	53.0
CNN+LSTM+SA	68.5	52.2	71.1	73.5	85.3	52.3
CNN+LSTM+SA*	76.6	64.4	82.7	77.4	82.6	75.4
CNN+LSTM+RN	95.5	90.1	97.8	93.6	97.9	97.1

* Our implementation, with optimized hyperparameters and trained fully end-to-end.

Table 1: **Results on CLEVR from pixels.** Performances of our model (RN) and previously reported models [16], measured as accuracy on the test set and broken down by question category.

fine-tuning, very large LSTMs for language encoding, and further processing modules, such as stacked or iterative attention, or large fully connected layers (upwards of 4000 units, often) [15].

5 Results

5.1 CLEVR from pixels

Our model achieved state-of-the-art performance on CLEVR at 95.5%, exceeding the best model trained only on the pixel images and questions at the time of the dataset’s publication by 27%, and surpassing human performance in the task (see Table 1 and Figure 3).

These results – in particular, those obtained in the **compare attribute** and **count** categories – are a testament to the ability of our model to do relational reasoning. In fact, it is in these categories that state-of-the-art models struggle most. Furthermore, the relative simplicity of the network components used in our model suggests that the difficulty of the CLEVR task lies in its relational reasoning demands, not on the language or the visual processing.

Results using privileged training information A more recent study reports overall performance of 96.9% on CLEVR, but uses additional supervisory signals on the functional programs used to generate the CLEVR questions [16]. It is not possible for us to directly compare this to our work since we do not use these additional supervision signals. Nonetheless, our approach greatly outperforms a version of their model that was not trained with these extra signals, and even versions of their model trained using 9K or 18K ground-truth programs. Thus, RNs can achieve very competitive, and even super-human results under much weaker and more natural assumptions, and even in situations when functional programs are unavailable.

5.2 CLEVR from state descriptions

To demonstrate that the RN is robust to the form of its input, we trained our model on the state description matrix version of the CLEVR dataset. The model achieved an accuracy of 96.4%. This result demonstrates the generality of the RN module, showing its capacity to learn and reason about object relations while being agnostic to the kind of inputs it receives – i.e., to the particular representation of the object features to which it has access. Therefore, RNs are not necessarily restricted to visual problems, and can thus be applied in very different contexts, and to different tasks that require relational reasoning.

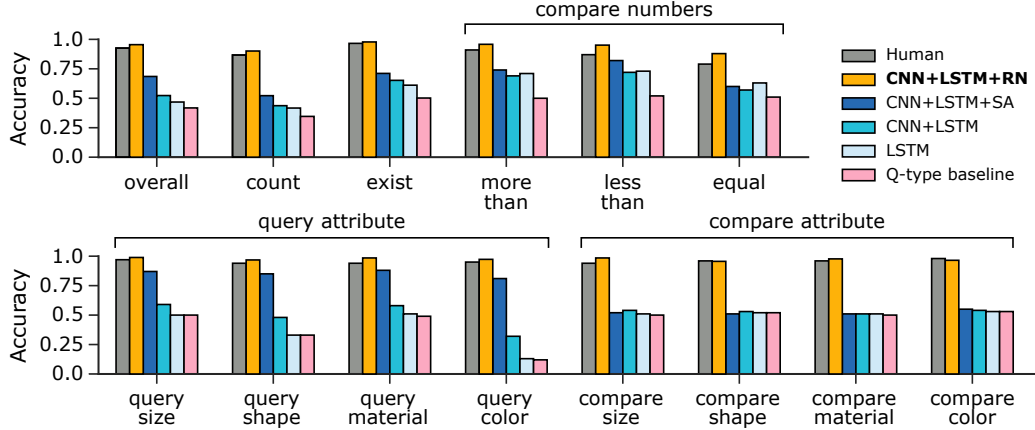


Figure 3: **Results on CLEVR from pixels.** The RN augmented model outperformed all other models and exhibited super-human performance overall. In particular, it solved “compare attribute” questions, which trouble all other models because they heavily depend on relational reasoning.

5.3 Sort-of-CLEVR from pixels

The results so far led us to hypothesize that the difficulty in solving CLEVR lies in its heavy emphasis on relational reasoning, contrary to previous claims that the difficulty lies in question parsing [17]. However, the questions in the CLEVR dataset are not categorized based on the degree to which they may be relational, making it hard to assess our hypothesis. Therefore, we use the Sort-of-CLEVR dataset which we explicitly designed to separate out relational and non-relational questions (see Section 3.2).

We find that a CNN augmented with an RN achieves an accuracy above 94% for both relational and non-relational questions. However, a CNN augmented with an MLP only reached this performance on the non-relational questions, plateauing at 63% on the relational questions. This strongly indicates that models lacking a dedicated relational reasoning component struggle, or may even be completely incapable of solving tasks that require very simple relational reasoning. Augmenting these models with a relational module, like the RN, is sufficient to overcome this hurdle.

A simple “closest-to” or “furthest-from” relation is particularly revealing of a CNN+MLP’s lack of general reasoning capabilities (52.3% success). For these relations a model must gauge the distances between *each* object, and then compare each of these distances. Moreover, depending on the images, the relevant distance could be quite small in magnitude, or quite large, further increasing the combinatoric difficulty of this task.

5.4 bAbI

Our model succeeded on 18/20 tasks. Notably, it succeeded on the basic induction task (2.1% total error), which proved difficult for the Sparse DNC (54%), DNC (55.1%), and EntNet (52.1%). Also, our model did not catastrophically fail in any of the tasks: for the 2 tasks that it failed (the “two supporting facts”, and “three supporting facts” tasks), it missed the 95% threshold by 3.1% and 11.5%, respectively. We also note that the model we evaluated was chosen based on overall performance on a withheld validation set, using a single seed. That is, we did not run multiple replicas with the best hyperparameter settings (as was done in other models, such as the Sparse DNC, which demonstrated performance fluctuations with a standard deviation of more than ± 3 tasks passed for the best choice of hyperparameters).

5.5 Dynamic physical systems

Finally, we trained our model on two tasks requiring reasoning about the dynamics of balls moving along a surface. In the connection inference task, our model correctly classified all the connections in

93% of the sample scenes in the test set. In the counting task, the RN achieved similar performance, reporting the correct number of connected systems for 95% of the test scene samples. In comparison, an MLP with comparable number of parameters was unable to perform better than chance for both tasks. Moreover, using this task to learn to infer relations results in transfer to unseen motion capture data, where RNs predict the connections between body joints of a walking human (see supplementary information for experimental details and example videos).

6 Discussion and Conclusions

This work showed how the RN, a dedicated module for computing inter-entity relations, can be plugged into broader deep learning architectures to substantially improve performance on tasks that demand rich relational reasoning. Our CLEVR results included super-human performance at 95.5% overall. Our bAbI results demonstrated broad reasoning capabilities, solving 18/20 tasks with no catastrophic failures. Together these results demonstrate the flexibility and power of this simple neural network building block.

One of the most interesting aspects of the work is that RN module inclusion in relatively simple CNN- and LSTM-based VQA architectures raised the performance on CLEVR from 68.5% to 95.5% and achieved state-of-the-art, super-human performance. **We speculate that the RN provided a more powerful mechanism for flexible relational reasoning, and freed up the CNN to focus more exclusively on processing local spatial structure.** This distinction between *processing* and *reasoning* is important. Powerful deep learning architectures, such as ResNets, are highly capable visual processors, but they may not be the most appropriate choice for reasoning about arbitrary relations.

A key contribution of this work is that the RN was able to induce, through the learning process, upstream processing to provide a set of useful object-like representations. Note, the input data and target objective functions did not specify any particular form or semantics of the internal object representations. This demonstrates the RN’s rich capacity for structured reasoning even with unstructured inputs and outputs.

Future work should apply RNs to a variety of problems that can benefit from structure learning and exploitation, such as rich scene understanding in RL agents, modeling social networks, and abstract problem solving. Future work could also improve the efficiency of RN computations. Though our results show that no knowledge about the particular relations among objects are necessary, RNs can exploit such knowledge if available or useful. For example, if two objects are known to have no actual relation, the RN’s computation of their relation can be omitted. An important direction is exercising this option in circumstances with strict computational constraints, where, for instance, attentional mechanisms could be used to filter unimportant relations and thus bound the otherwise quadratic complexity of the number of considered pairwise relations.

Relation Networks are a simple and powerful approach for learning to perform rich, structured reasoning in complex, real-world domains.

Acknowledgments

We would like to thank Murray Shanahan, Ari Morcos, Scott Reed, Daan Wierstra, Alex Lerchner, and many others on the DeepMind team, for critical feedback and discussions.

Supplementary Material

Here we provide additional details on (A) related work, (B) CLEVR from pixels, (C) CLEVR from state descriptions, (D) Sort-of-CLEVR, (E) bAbI, and (F) Dynamic physical system reasoning. For each task, we provide additional information on the dataset, model architecture, training and results where necessary.

A Related Work

Since the RN is highly versatile, it can be used for visual, text-based, and state-based tasks. As such, it touches upon a broad range of areas in machine learning, computer vision, and natural language understanding. Here, we provide a brief overview of some of the most relevant related work.

Relational reasoning

Relational reasoning is implicit in many symbolic approaches [11, 32] and has been explicitly pursued using neural networks as well [4]. There is recent work applying neural networks to graphs, which are a natural structure for formalising relations [12, 19, 33, 37, 26, 2]. Perhaps a crucial difference between this work and our work here is that RNs require minimal oversight to produce their input (a set of objects), and can be applied successfully to tasks even when provided with relatively unstructured inputs coming from CNNs and LSTMs. There has also been some recent work on reasoning about sets, although this work does not explicitly reason about the *relations* of elements *within* sets [47].

Grounding spatial relations

Although grounding language in spatial percepts has a long-standing tradition, the majority of previous research has focused on either rule-based spatial representations or hand-engineered spatial features [8, 10, 20, 21, 24, 29, 38, 39]. Although there are some attempts to learn spatial relations using spatial templates [28, 30], these approaches are less versatile than ours.

Visual question answering

Visual question answering is a recently introduced task that measures a machine understanding of the scene through questions [1, 29]. Related to our work, we are mostly interested in the newly introduced CLEVR dataset [15] that distills core challenges of the task, namely relational and multi-modal reasoning. The majority of approaches to question answering share the same pipeline [6, 31, 36]. First, questions are encoded with recurrent neural networks, and images are encoded with convolutional neural networks. Next, both representations are combined, and the answers are either predicted or generated. Most successful methods also use an attention mechanism that locate important image regions [5, 44, 45, 46]. In our work, we follow a similar pipeline, but we use Relation Networks as a powerful reasoning module.

Parallel to our work, two architectures have shown impressive results on the CLEVR dataset [14, 16]. Both approaches hinge on compositionality principles, and have shown they are capable of some relational reasoning. However, both require either designing modules, or require direct access to ground-truth programs. The RN module, on the other hand, is conceptually simpler, can readily be combined with basic neural components such as CNNs or LSTMs, can be broadly applied to various tasks, and achieves significantly better results on CLEVR [15] than [14], and on par with strongly supervised system of [16].

Text-based question answering

Answering text-based questions has long been an active research area in the NLP community [3, 22, 27, 48]. Recently, in addition to traditional symbolic-based question answering architectures, we observe a growing interest in neural-based approaches to text based question answering [34, 42, 43]. While these architectures rely on ‘memories’, we empirically show that the RN module has similar

capabilities, reaching very competitive results on the bAbI dataset [41] – a dataset that tests reasoning capabilities of text-based question answering models.

B CLEVR from pixels

Our model (described in Section 4 of the main text) was trained on 70000 scenes from the CLEVR dataset and a total of 699989 questions. Images were first down-sampled to size 128×128 , then pre-processed with padding to size 136×136 , followed by random cropping back to size 128×128 and slight random rotations between -0.05 and 0.05 rads. We used 10 distributed workers that synchronously updated a central parameter server. Each worker learned with mini-batches of size 64, using the Adam optimizer and a learning rate of $2.5e^{-4}$. Dropout of 50% was used on the penultimate layer of the RN. In our best performing model each convolutional layer used 24 kernels of size 3×3 and stride 2, batch normalization, and rectified linear units. The model stopped improving in performance after approximately 1.4 million iterations, at which point training was concluded. The model achieved 96.8% accuracy on the validation set. In general, we found that smaller models performed best. For example, 128 hidden unit LSTMs performed better than 256 or 512, and CNNs with 24 kernels were better than CNNs with more kernels, such as 32, 64, or more.

Failure cases

Although our model gets most answers correct, a closer examination of the failure cases help us to identify limitations of our architecture. In Table 2, we show some examples of CLEVR questions that our model fails to answer correctly, along with the ground-truth answers. Based on our observations, we hypothesize that our architecture fails especially when objects are heavily occluded, or whenever a high precision object position representation is required. We also observe that many failure cases for our model are also challenging for humans.

C CLEVR from state descriptions

The model that we train on the state description version of CLEVR is similar to the model trained on the pixel version of CLEVR, but without the vision processing module. We used a 256 unit LSTM for question processing and word-lookup embeddings of size 32. For the RN we used a four-layer MLP with 512 units per layer, with ReLU non-linearities for g_θ . A three-layer MLP consisting of 512, 1024 (with 2% dropout) and 29 units with ReLU non-linearities was used for f_θ . To train the model we used 10 distributed workers that synchronously updated a central parameter server. Each worker learned with mini-batches of size 64, using the Adam optimizer and a learning rate of $1e^{-4}$.

D Sort-of-CLEVR

The Sort-of-CLEVR dataset contains 10000 images of size 75×75 , 200 of which were withheld for validation. There were 20 questions generated per image (10 relational and 10 non-relational).

Non-relational questions are split into three categories: (i) query shape, e.g. “*What is the shape of the red object?*”; (ii) query horizontal position, e.g. “*Is the red object on the left or right of the image?*”; (iii) query vertical position, e.g. “*Is the red object on the top or bottom of the image?*”. These questions are non-relational because one can answer them by reasoning about the attributes (e.g. position, shape) of a single entity which is identified by its unique color (e.g. *red*).

Relational questions are split into three categories: (i) closest-to, e.g. “*What is the shape of the object that is closest to the green object?*”; (ii) furthest-from, e.g. “*What is the shape of the object that is furthest from the green object?*”; (iii) count, e.g. “*How many objects have the shape of the green object?*”. We consider these relational because answering them requires reasoning about the attributes of one or more objects that are defined relative to the attributes of a reference object. This reference object is uniquely identified by its color.

Questions were encoded as binary strings of length 11, where the first 6 bits identified the color of the object to which the question referred, as a one-hot vector, and the last 5 bits identified the question type and subtype.

In this task our model used: four convolutional layers with 32, 64, 128 and 256 kernels, ReLU non-linearities, and batch normalization; the questions, which were encoded as fixed-length binary strings, were treated as question embeddings and passed directly to the RN alongside the object pairs; a four-layer MLP consisting of 2000 units per layer with ReLU non-linearities was used for g_θ ; and a four-layer MLP consisting of 2000, 1000, 500, and 100 units with ReLU non-linearities used for f_ϕ . An additional final linear layer produced logits for a softmax over the possible answers. The softmax output was optimized with a cross-entropy loss function using the Adam optimizer with a learning rate of $1e^{-4}$ and mini-batches of size 64.

We also trained a comparable MLP based model (CNN+MLP model) on the Sort-of-CLEVR task, to explore the extent to which a standard model can learn to answer relational questions. We used the same CNN and LSTM, trained end-to-end, as described above. However, this time we replaced the RN with an MLP with the same number of layers and number of units per layer. Note that there are more parameters in this model because the input layer of the MLP connects to the full CNN image embedding.

E bAbI model for language understanding

For the bAbI task, each of the 20 sentences in the support set was processed through a 32 unit LSTM to produce an object. For the RN, g_θ was a four-layer MLP consisting of 256 units per layer. For f_ϕ , we used a three-layer MLP consisting of 256, 512, and 159 units, where the final layer was a linear layer that produced logits for a softmax over the answer vocabulary. A separate LSTM with 32 units was used to process the question. The softmax output was optimized with a cross-entropy loss function using the Adam optimizer with a learning rate of $2e^{-4}$.

F Dynamic physical system reasoning

For the connection inference task the targets were binary vectors representing the existence (or non-existence) of a connection between each ball pair. For a total of 10 objects, the targets were 10^2 length vectors. For the counting task, the targets were one-hot vectors (of length 10) indicating the number of systems of connected balls. It is important to point out that in the first task the supervision signal provided by the targets explicitly informs about the relations that need to be computed. In the second task, the supervision signal (counts of systems) do not provide explicit information about the kind of relations that need to be computed. Therefore, the models that solve the counting task must successfully infer the relations implicitly.

Inputs to the RN were state descriptions. Each row of a state description matrix provided information about a particular object (i.e. ball), including its coordinate position and color. Since the system was dynamic, and hence evolved through time, each row contained object property descriptions for 16 consecutive time-frames. For example, a row could be comprised of 33 floats: 16 for the object's x coordinate position across 16 frames, 16 for the object's y coordinate position across 16 frames, and 1 for the object's color. The RN treated each row in this state description matrix as an object. Thus, it had to infer an object description contained information of the object's properties evolving through time.

For the connection inference task, the RN's g_θ was a four-layer MLP consisting of three layers with 1000 units and one layer with 500 units. For f_ϕ , we used a three-layer MLP consisting of 500, 100, and 100 units, where the final layer was a linear layer that produced logits corresponding to the existence/absence of a connection between each ball pair. The output was optimized with a cross-entropy loss function using the Adam optimizer with a learning rate of $1e^{-4}$ and a batch size of 50. The same model was used for the counting task, but this time the output layer of the RN was a linear layer with 10 units. For baseline comparisons we replaced the RNs with MLPs with comparable number of parameters.

Please see the supplementary videos:

<https://www.youtube.com/channel/UCIAAnkrNn45D0MeYwtVpmbUQ>

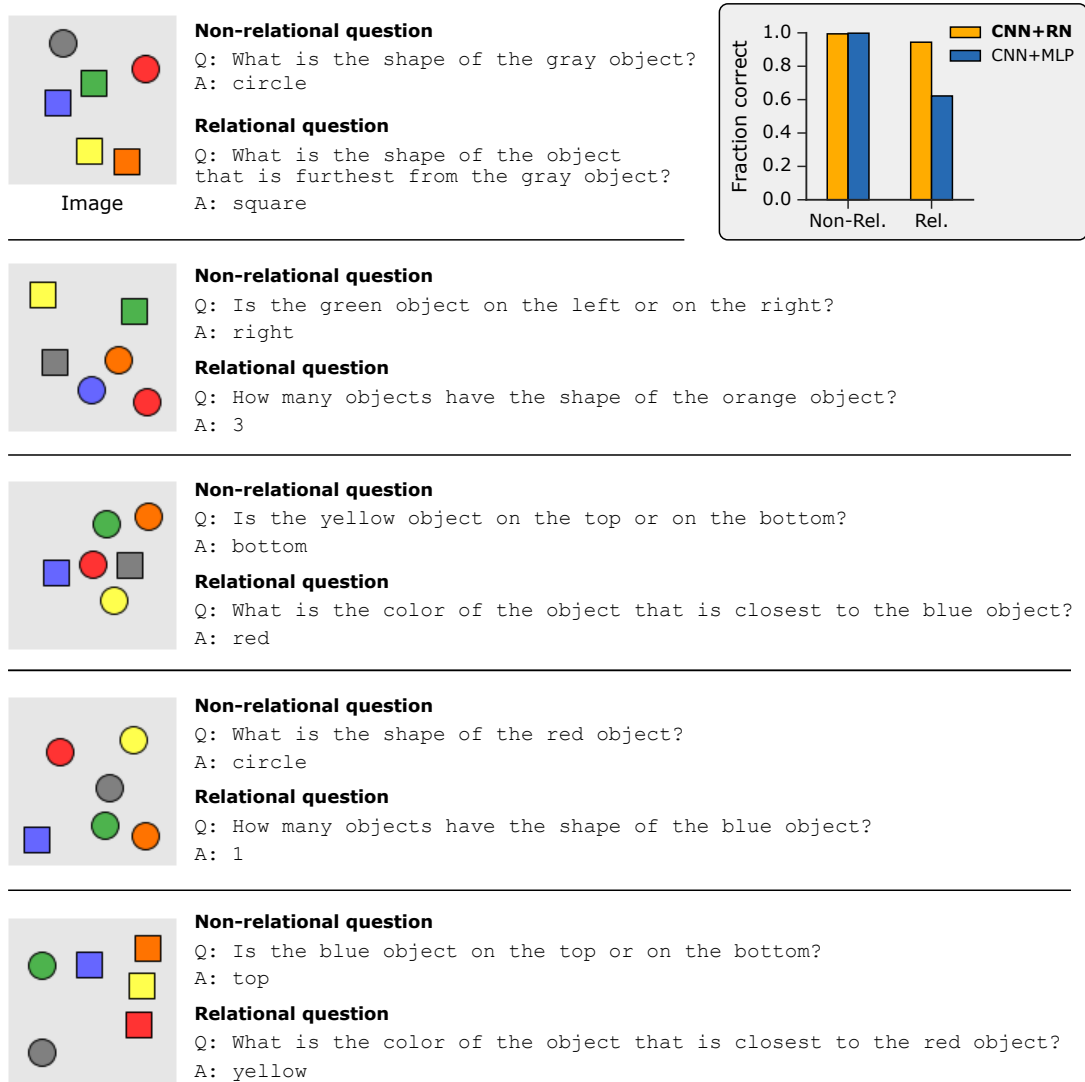


Figure 4: “Sort-of-CLEVR” task: examples and results. The Sort-of-CLEVR example here consists of an image of six objects and two questions – a relational question, and a non-relational question – along with the corresponding answers. The fraction of correctly answered relational questions (inset bar plot) for our model (CNN+RN) is much larger than the comparable MLP based model (CNN+MLP), whereas both models have similar performance levels for non-relational questions.

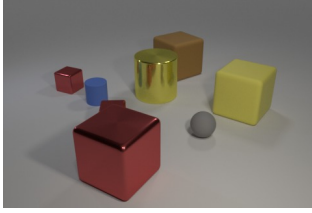
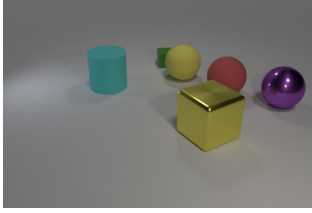
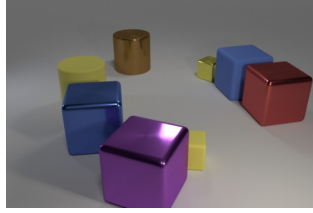
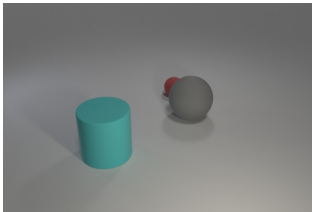
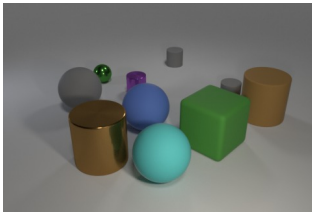
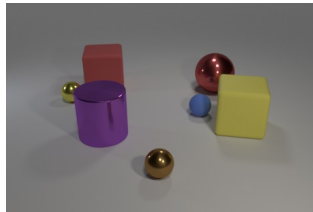
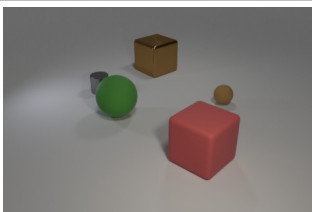
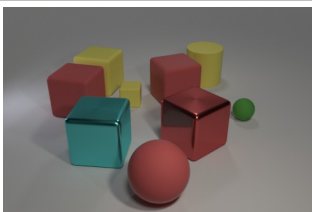
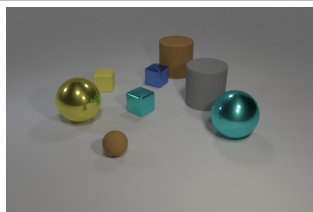
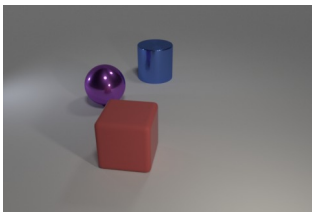
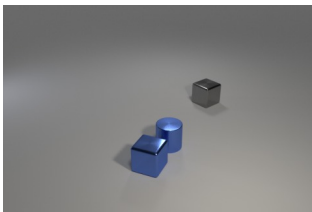
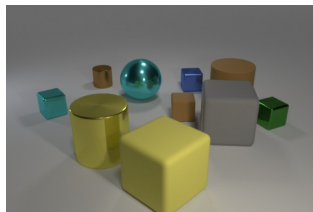
		
What shape is the small object that is in front of the yellow matte thing and behind the gray sphere?	What number of things are either tiny green rubber objects or shiny things that are behind the big metal block?	What number of objects are blocks that are in front of the large red cube or green balls?
<i>RN:</i> cylinder	1	2
<i>GT:</i> cube	2	3
		
Is the shape of the small red object the same as the large matte object that is right of the small rubber ball?	How many gray objects are in front of the tiny green shiny ball and right of the big blue matte thing?	What number of objects are big red matte cubes or things on the right side of the large red matte block?
<i>RN:</i> no	0	5
<i>GT:</i> yes	1	6
		
There is a brown ball; what number of things are left of it?	How many objects are big purple rubber blocks or red blocks in front of the tiny yellow rubber thing?	How many things are rubber cylinders in front of the tiny yellow block or blocks that are to the right of the small brown rubber thing?
<i>RN:</i> 3	3	2
<i>GT:</i> 4	2	3
		
What number of objects are either big things that are left of the cylinder or cylinders?	Are there the same number of small blue objects that are to the right of the blue cube and blue metal cubes?	What number of other things are there of the same material as the green cube?
<i>RN:</i> 2	no	6
<i>GT:</i> 3	yes	5

Table 2: Failures on CLEVR; RN – predicted answers, GT – ground-truth answer.

References

- [1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015.
- [2] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, et al. Interaction networks for learning about objects, relations and physics. In *NIPS*, 2016.
- [3] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, 2013.
- [4] Rajarshi Das, Arvind Neelakantan, David Belanger, and Andrew McCallum. Chains of reasoning over entities, relations, and text using recurrent neural networks. *arXiv:1607.01426*, 2016.
- [5] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv:1606.01847*, 2016.
- [6] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. In *NIPS*, 2015.
- [7] Marta Garnelo, Kai Arulkumaran, and Murray Shanahan. Towards deep symbolic reinforcement learning. *arXiv:1609.05518*, 2016.
- [8] Dave Golland, Percy Liang, and Dan Klein. A game-theoretic approach to generating spatial descriptions. In *EMNLP*, 2010.
- [9] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 2016.
- [10] Sergio Guadarrama, Lorenzo Riano, Dave Golland, Daniel Gouhring, Yangqing Jia, Dan Klein, Pieter Abbeel, and Trevor Darrell. Grounding spatial relations for human-robot interaction. In *IROS*, 2013.
- [11] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.
- [12] Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv:1506.05163*, 2015.
- [13] Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. Tracking the world state with recurrent entity networks. In *ICLR*, 2017.
- [14] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. *arXiv:1704.05526*, 2017.
- [15] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.
- [16] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. *arXiv:1705.03633*, 2017.
- [17] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. *arXiv:1703.09684*, 2017.
- [18] Charles Kemp and Joshua B Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, 2008.
- [19] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv:1609.02907*, 2016.
- [20] Jayant Krishnamurthy and Thomas Kollar. Jointly learning to parse and perceive: Connecting natural language to the physical world. *TACL*, 2013.
- [21] Geert-Jan M Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I Christensen. Situated dialogue and spatial organization: What, where... and why. *IJARS*, 2007.
- [22] Tom Kwiatkowski, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. Inducing probabilistic ccg grammars from logical form with higher-order unification. In *EMNLP*, 2010.
- [23] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *arXiv:1604.00289*, 2016.

- [24] Tian Lan, Weilong Yang, Yang Wang, and Greg Mori. Image retrieval with structured object queries using latent ranking svm. In *ECCV*, 2012.
- [25] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [26] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *ICLR*, 2016.
- [27] Percy Liang, Michael I Jordan, and Dan Klein. Learning dependency-based compositional semantics. *Computational Linguistics*, 2013.
- [28] Gordon D Logan and Daniel D Sadler. A computational analysis of the apprehension of spatial relations. 1996.
- [29] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, 2014.
- [30] Mateusz Malinowski and Mario Fritz. A pooling approach to modelling spatial relations for image retrieval and annotation. *arXiv:1411.5190*, 2014.
- [31] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A deep learning approach to visual question answering. *arXiv:1605.02697*, 2016.
- [32] Allen Newell. Physical symbol systems. *Cognitive science*, 4(2):135–183, 1980.
- [33] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *ICML*, 2016.
- [34] Jack Rae, Jonathan J Hunt, Ivo Danihelka, Timothy Harley, Andrew W Senior, Gregory Wayne, Alex Graves, and Tim Lillicrap. Scaling memory-augmented neural networks with sparse reads and writes. In *NIPS*, 2016.
- [35] David Raposo, Adam Santoro, David Barrett, Razvan Pascanu, Timothy Lillicrap, and Peter Battaglia. Discovering objects and their relations from entangled scene representations. *arXiv:1702.05068*, 2017.
- [36] Mengye Ren, Ryan Kiros, and Richard Zemel. Image question answering: A visual semantic embedding model and a new dataset. In *NIPS*, 2015.
- [37] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 2009.
- [38] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth J Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*, 2011.
- [39] Stefanie Tellex, Thomas Kollar, George Shaw, Nicholas Roy, and Deb Roy. Grounding spatial language for video search. In *ICMI*, 2010.
- [40] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IROS*, 2012.
- [41] Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv:1502.05698*, 2015.
- [42] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *ICLR*, 2015.
- [43] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *ICML*, 2016.
- [44] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, 2016.
- [45] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.
- [46] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.
- [47] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. *arXiv:1703.06114*, 2017.
- [48] John M Zelle and Raymond J Mooney. Learning to parse database queries using inductive logic programming. In *AAAI*, 1996.