Dissertation Type: software development

DEPARTMENT OF COMPUTER SCIENCE

Incorporating Relational Reasoning into Neural Network

Lizhao Liu

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering.

Monday 13th September, 2021

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of MSc in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Lizhao Liu, Monday 13th September, 2021

Abstract

This projects aims to incorporating relational reasoning ability into a neural network in order to handle visual reasoning tasks inspired by the RFT ideas. To this end, this project adapted the Relation Networks as the base reasoning module with its' relation-centric design and superb relational reasoning ability. Our neural network model is tested with the Natural Language for Visual Reasoning dataset, which requires the model to reason out the spatial, quantitative or comparative relations between the shapes from the visual inputs. Due to the scalability problem and the task complexity limitation of the basic model, we also implemented a enhanced version model combined an attention mechanism and a double-layer design. Our model finally achieves 55.7% accuracy for the first version and 60.7% accuracy for the enhanced version, which achieves a comparably high performance according to the leaderboard on the official NLVR website, which contributes to the development of enabling AI to simulate human's brain underlying process in cognition building and language learning.

key words: Relational Frame Theory, Relation Networks, visual reasoning tasks, language learning

Contents

1 Introduction

1.1 Motivation

The most central component of human intelligence is language and cognition [6]. As human beings, we are endowed with great talent of talking, understanding and reasoning. All of these contribute to the development of our fully functional communication in social society [24], thus leading to the advancement and blossom in human civilisation, science, literature and economy. Therefore, understanding and accounting for the process that underlies human cognition and communication is critical though fraught with difficulty and remains a challenge for psychologists to overcome[6].

In the recent years, the Relational Frame Theory (RFT) was established as a comprehensive account of human language and cognition ability, which embraces the simple idea that such ability lies in our unique and great relating ability [12, 9, 6]. In other words, humans are able to identify and create associations between stimuli such as words objects and events [2], which enables our brain to be active and to produce or understand completely novel sentences, thus achieving successful communication with others under social context [24].

Therefore, how can we make the AI intelligence to master such ability to reason our objects and their interactions, and simulate humans' language learning process has attracted many AI partitioners in the recent years, which becomes the focus of this project. In 2017, Relation Networks (RNs) proposed by Santoro et al. [22] can be seen as a milestone in this field [16]. RNs are developed as a neural network module which is fundamentally relation-centric. RNs take relations as explicit inputs of a neural network instead of individual object, thus allowing all of these potential interactions between objects to be evaluated during the training process [22]. Therefore, we decided to choose RNs as our base model as it, similar to RFT ideas, treats everything as stimuli of a relation and is able to infer both physical and abstract relations from the

learning (training) process. Such similarities from the underlying processes of RNs and our active human brain makes RNs a the suitable choice to achieve a powerful neural network with powerful relational reasoning capacity.

1.2 Aim and Objectives

The aim of this project is to implement the Relational Frame Theory and simulate humans' relational reasoning ability through a neural network for visual reasoning tasks. In this way, we attempt to enable the neural network to simulate humans' language learning process, master the ability to understand novel sentences and give correct response, thereby contributing to AI's development of cognition, reasoning and language ability.

In order to achieve this aim successfully, we sets up the following objectives:

1. Understanding the Relational Frame Theory (RFT) and the underlying foundation of human recognition and communication ability from the perspective of RFT.
2. Incorporating the Relation Networks module into a neural network to simulate RFT ideas and test the model's performance with Natural Language for Visual Reasoning (NLVR) dataset.
3. Understanding the limitations and current enhancements of the original Relation Networks model.
4. Combining the features from other RNs variations and build the enhanced model for NNLVR tasks.
5. Evaluating the enhancement features of the RNs model.
6. Evaluating the RNs approach with comparison of other current methods being used in the NLVR dataset.

1.3 Thesis Layout

The thesis is structured as follows: Chapter 2 introduces background of this project, which starts with an overview of the proposals and features of Relational Frame Theory and goes on to the introduction of Relational Networks approach and the reasons why we choose RNs as our base model. Chapter 3 describes the implementation process of our project, which includes the basic model's structure and details, the experiment carried on the basic model, the discoveries and limitations of the basic model's reasoning performance, the review of other current enhancement of the RNs model, and our enhanced version of the RNs model based on the combination of other RNs variations. Chapter 4 presents the results obtained from applying NRs models to NLVR dataset, and discusses about the improvements from the basic model to the enhanced model, the comparison with other approaches and the limitations of our approach shown from the result. Chapter 5 concludes a the results of the project, the contributions the project made and suggests the future work.

2 Background

2.1 Relational Frame Theory

2.1.1 Human Language and Cognition

It is generally believed that human surpass other species as we have the unique ability to generate language. Stewart [24] defines the ability to generate language as the ability to produce or understand completely novel sentences that never said or heard before, which typically involves giving a response without being trained. For example, a child is taught with the word "dog" when seeing a dog of any kind. Then, he will be able to say the word "dog" upon seeing or hearing an animal with the dogs' features like four

legs, one tail or a barking voice, even if this dog is never shown to him before. Furthermore, he also points out such response generalization is considered as fundamental to human cognition and language, which in a way contributes to the development of fully functional communication in social society [24]. Therefore, analyzing and accounting for the process that underlies human cognition and communication is critical and has been the working direction that psychologists or behaviorists are devoted to figure out [6].

2.1.2 Derived Relational Responding

In order to understand the underlying process of human cognition and communication, psychologists proposed the Relational Frame Theory (RFT), which can be seen as a behavior-analytic account of human language and cognition [12, 6]. From the perspective of RFT, the superb cognition and communication ability of human beings are founded in our powerful relating ability. Such relating ability enables us to identify, create or derive associations between stimuli like words, events or ideas [2]. For example, if we present a child a picture of a candy with the written word "candy", he will acquire the relationship between the candy picture and its' written form, thus being able to match the correct image with the textual input "candy". Then, we pronounce the word "candy" when showing the candy picture again. Later, he will be able to produce the appropriate spoken word "candy" in the presence of the textual word "candy". In this case, the child learns the relations "A equals to B", "B equals to C", and then he easily identifies the relation "A is equal to C". Despite such stimulus equivalence, human beings are also able to recognize relations such as opposition, comparison, distinction, analogy, time etc, all of which rely on our essential relational ability [24].

2.1.3 Properties of RFT

According to RFT, there are three properties of relational frames: mutual entailment, combinatorial entailment and transformation of function.

Mutual entailment involves two stimuli. It refers to relations that are bi-directional. For example, if stimulus A is related to stimulus B, stimulus B is believed to have a relationship with stimulus A. Such relations not only refer to symmetry which means the two stimulus are equivalent, but also refer to non-equivalent relations that are bi-directional.[9, 12]

Combinatorial entailment describes relations that are based on a combination of established relations, which normally refers to 3 or more stimuli. For example, it is easily to infer the relation "A is more than C" based on the combined relations "A is more than B" and "B is more than C". Such transfer of relations or transitivity contribute to our ability to infer complex relations in real world.[9, 12]

Transformation of function refers to a change of the existing stimulus functions caused by their change of participation in relational frame. For example, if a child relates a scared feeling with dogs, after having a lovely puppy as a pet and spending some happy time with it, the relationship will be changed and the child will associate happy mood with dogs.[9, 2]

2.1.4 Human's Relating Ability vs. Other Species' Relating Ability

Compared with other species, out relating ability surpass in the following aspects:

1. Human beings are able to make associations not only based on the obvious feature of the link, but also find connections from other dimensions that come along with

the relevant stimuli. For example, a lamp can be associated with the sun as they both light the dark place. As human beings, we are able to recognize more types and dimensions of their relation as same function, same colour, different size, is or not touchable etc.

2. Other species can only make simple physical associations, but human beings are able to recognize different types of abstract relations such as comparison, distinction, opposition, analogy, time, difference and so on between stimuli, thus forming a relation framework.

3. Humans are able to derive complex relational networks under contextual and social cues. During the language learning process, it is believed that humans are able to derive relations from the contextual or arbitrary cues, rather than the physical and formal attributes presented between stimuli. For example, if the context indicates the relation "A is bigger than B", humans can obtain this relationship as if A is actually bigger than B. In addition, humans will also be able to respond to B as if it is actually smaller than A according to mutual entailment from the perspective of RFT. Similarly, with the feature of combinatorial entailment, one will also be able to identify all of the relations between A, B and C after being taught "A is bigger than B" and "B is bigger than C". All of these relations are arbitrary without referring to the physical properties of these stimuli and some are not directly trained to humans by virtue of combinatorial or mutual entailment of humans' relating frame, thus forming our unique and superb relating skills with our "arbitrarily applicable relational responding" ability.[12]

2.2 Relation Networks (RNs)

2.2.1 Approaches to Relational Reasoning

From the perspective of RFT, it is concluded that the relating ability is a central component of human intelligence, which contributes to our ability to generate novel language

and response, thus forming the foundation of all human language. However, replicating such relational reasoning ability of humans on artificial intelligence systems remains a key challenge. In the recent years, many approaches have been proposed to simulate human's complex relational reasoning ability in neural networks. Symbolic approaches are generally used to complex reasoning tasks, which use mathematical or logical methods to identify the possible relations between symbols and infer further relations with a variety of methods such as deduction, and induction [22]. However, such approach suffers from the symbol grounding problem [14] and cannot handle unstructured inputs well. Other methods such as statistical approach, try to construct representations from raw inputs and identify relations, but is at a disadvantage with data-poor problems, where a complex but sparse underlying relationships lies behind the structure.[22]

2.2.2 Relation Networks

Relation Networks (RNs) is a neural network module primarily designed with a structure focusing on relational reasoning. Compared with other traditional neural networks, RNs are fundamentally relation-centric with great flexibility and applicability. Instead of processing individual objects as inputs from previous layers, RNs take pairwise relations between objects as explicit inputs, thereby allowing the module to process all potential interactions with considerations of all combinations of input objects. [22] Also, the input objects are expected to contain all necessary attributes that are necessarily needed to reason out their relations [16].

Let us consider a task which requires a neural network to give the correct response when it is required to judge the truth value of a statement "the mouse will run away" upon seeing an image of a mouse and a cat bed. With the traditional neural network approaches, the neural network will provide the correct answer "true" after a certain period of training process. However, the model gives the appropriate answer as the training dataset told it. The model knows nothing about the relations among the input objects "a mouse", "a cat", and "a cat bed". However, for Relation Networks, each

of these pairwise objects' relations will be passed to the neural network module and evaluated by the model (as figure 1 shows) [16]. In this way, the model is able to reason out how each of these two objects are related, thus giving the correct answer "true" for the statement "the mouse will run away" based on the relations "the mouse will run away when seeing a cat", and "a cat sleeps on a cat bed".

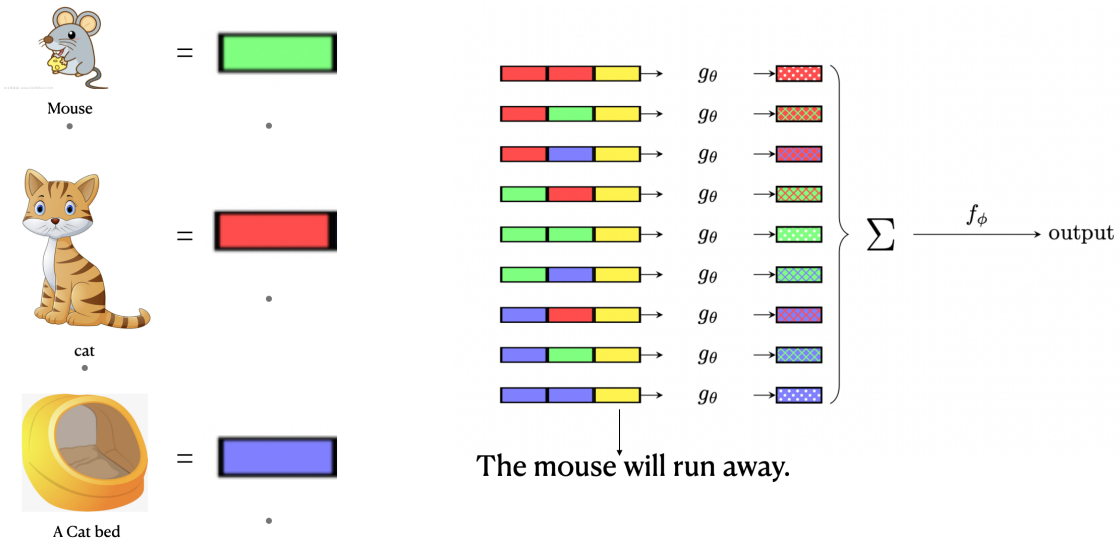


Figure 1: How Relation Networks Works?

2.2.3 Advantages of Relation Networks

There are a number of neural network approaches with a focus on relation-centric computation, such as Interaction Networks [7] and Graph Neural Networks [23]. However, RNs distinguish with other approaches with its flexibility, and similarity to the RFT ideas. Therefore, to the end of our project to simulate human brain's powerful complex relational reasoning ability in neural networks from the perspective of RFT, we decides to use RNs as our base model based on the following considerations:

1. The flexibility of RNs enables itself to reason out different relations from unstructured data. The RN use MLPs to detect objects from the raw or perceptual inputs and evaluate all combinations of input objects with regard to the corresponding

question/statement, which means the whole model will be trained end-to-end. Therefore, RNs are able to recognize appropriate object representations from a variety of raw and unstructured data inputs and study the various internal relations among different dataset accordingly, which simulates humans' everyday relating activities from a huge amount of various stimuli in the real world.[16]

2. Every object being processed through RNs will contain all attributes and information, both abstract and physical, which are necessary to reason out their relations. In this way, RNs are able to catch all types and dimensions of relations in the given dataset, thus simulating human's comprehensive relating ability to capture relations from various dimensions as the RFT ideas point out.

3. RNs are developed to consider the potential relations between all object pairs [22], which is similar to RFT ideas that the active human brain that treats everything as stimuli.

4. RNs feature with the ability to infer the relations from the training process. RNs take pairwise relations from all objects as inputs, and evaluate these relations using MLPs , which means it is not necessary for RNs to know which object actually exist, or the meaning of any particular relationship beforehand.[22] RNs are able to infer and learn the meaning and existence of relations during the training process, which also simulates RFT's proposals of human cognition and language generation process to great extent.

Therefore, we decided to build our model based on RNs as a way to simulate humans' powerful relating ability from the perspective of RFT ideas.

3 Implementation

3.1 Natural Language for Visual Reasoning Dataset

To test the model's ability on language learning and relational reasoning, we decided to run our model in a visual question answering dataset called Cornell Natural Language for Visual Reasoning (NLVR) dataset [25]. The NLVR dataset is composed of an image containing three boxes with different objects and one sentence which is grounded in the image. The task is to judge whether the truth value of a statement based on the visual input. The sentences are natural human-written sentences collected through crowdsourcing. Solving the task entails the ability to reason about sets of objects in the image, comparing their properties, evaluating their spatial or quantitative relations etc. Each object are designed to contain four properties: position, color, shape and size[25] in this dataset. Figure 2 shows two examples from the dataset:

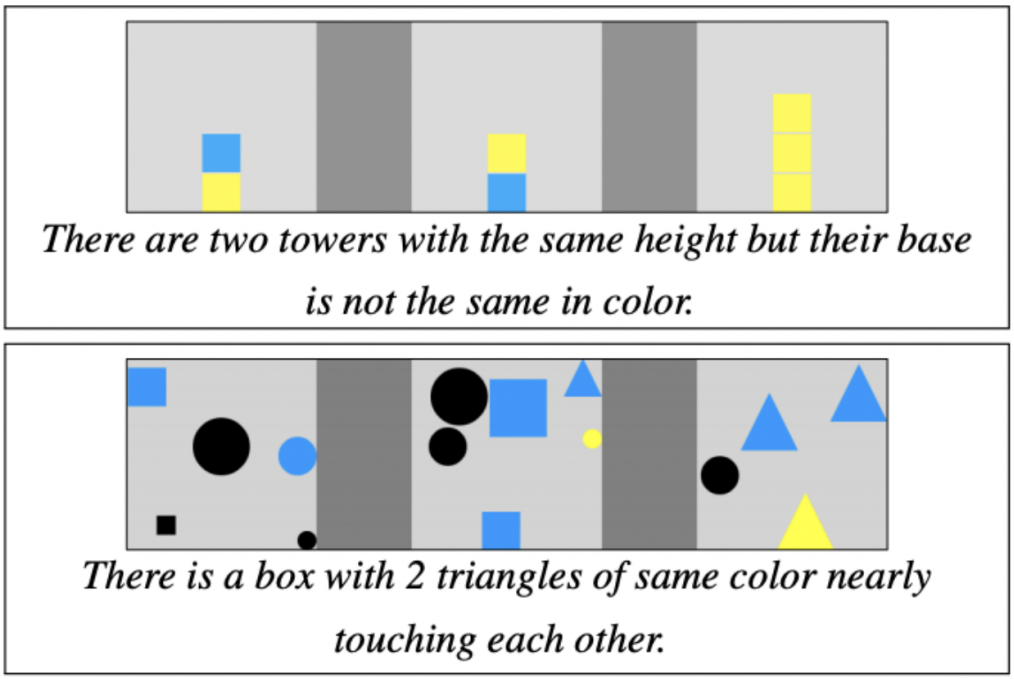


Figure 2: Examples from the NLVR dataset. Each image contains three subimages with different shapes. The judgement for the first statement is true. The second is false. [25]

We choose this dataset based on the following reasons:

1. Compared with high-level scene understanding visual QA dataset which requires knowledge that exceeds the information contained in the visual inputs or language inputs such as human common science or complicated knowledge of the world, or requires reasoning without fully specified language inputs, NLVR dataset presents the model with visual or language inputs that are sufficient for the reasoning tasks. In this way, the dataset focuses on the model's relational reasoning ability without influence by ambiguities, linguistic biases or other unavailable knowledge that are not contained in the training data[25].

2. The statement to be judged are written by real humans. Other similar visual answering dataset focusing on relational reasoning tasks like SHAPES [4] and

CLEVR [17] generate the question/statement with an automatic language generating mechanism grounded in the corresponding visual inputs. The NLVR dataset, instead, collects the statements through crowdsourcing process, which not only simulates a real language learning process with rich and unpredictable language input for a child to learn and understand, but also avoids the model to learn and interpret the task generating mechanism and requires the model to make predictions in an end-to-end manner.[25]

3. The task is designed to test the complex reasoning ability for challenges in visual and set-theoretic reasoning. During the crowdsourcing process, the workers are discouraged to make simple statements about the visual input [25]. In the example above, one statement will contain several nested types of relations about the properties of objects in image. Judging the truth value of such statement requires complex relational reasoning ability about the objects in the corresponding image. Therefore, NLVR serves as a suitable dataset to test the model's performance and relating capacity with comparison to real human brains.[25]

3.2 Implementation of the Basic Model

3.2.1 Model Design

Our model uses the Relation Networks [22] as the reasoning module. The whole system can be divided into 2 modules: input module and reasoning module (shown in figure 3), which adapts the ideas from [22]. The input information will go through these modules through a number of passes or hops and the output of the previous pass will adjust the current state, thereby performing some incremental refinement.

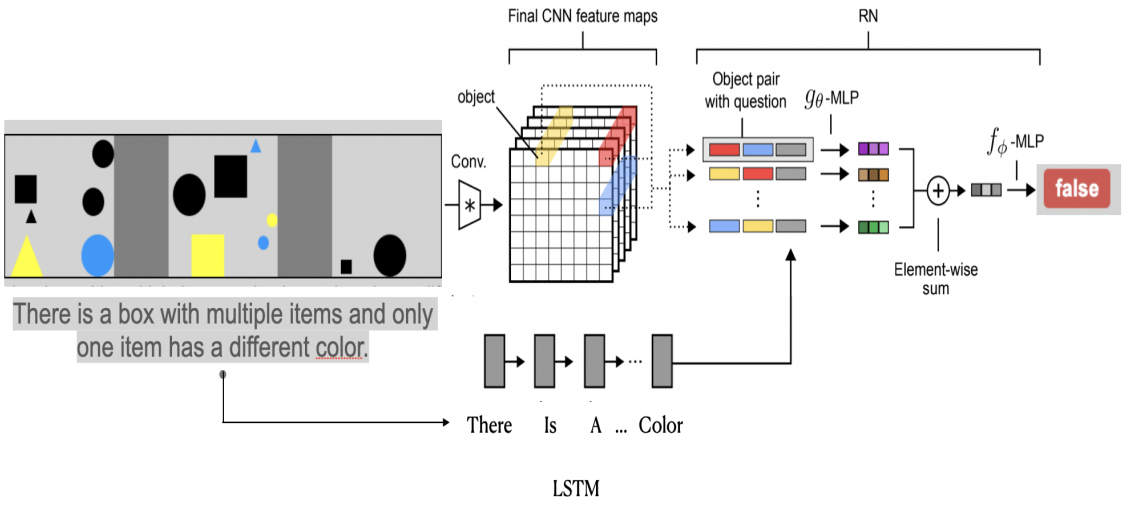


Figure 3: The RNs model applied to Natural Language for Visual Reasoning tasks. The figure is adapted from [22].

Input module The input module is designed to convert the raw information into internal object representations. In order to process the images, we used a Convolutional Neural Network (CNN) proposed by [22] in order to process the pixel inputs of the images into objects that contains information about their own properties. The CNN takes the images as input and convolves them through three convolutional layers and finally outputs k feature maps of size dd, where k shows "the number of kernels in the final convolutional layer" [22]. And every dd k-dimensional cell generated through the convolution process is seen as an object in this case. These objects could be of any form or type, such as a part of the background, a specific shape etc., which renders the model great flexibility [22]. In addition, in the NLVR dataset, the statements serve as the source of the meaning of an object-object relation. Therefore, to solve the task, the relations to be sought and attended should be statement-dependent. For instance, if the statement requires evaluations about triangles, then the relations between other shapes are irrelevant in this case. Therefore, we need to refer to the statement in question to condition the processing. To encode the statement, we adapted the method from Pavez

et al.[20] which used a word embedding projection method. We encoded each word in the statement into a vector v_i with the use of an embedding matrix. Then, we used a gated recurrent network [8] to encode the whole statement into a statement embedding vector q: $q = GRU([v_1, v_2, ..., v_M])$.

Reasoning module The reasoning module is designed to use the representations from the input module to reason about their interactions. We implemented the RNs module as the reasoning module [22]. The visual object vectors outputted from CNN are concatenated in pairs together with the encoded statement vector q. Then, each pair is then passed to a three-layer perceptron neural network g and all the outputs generated are summed into a single vector and then passed to a final neural network f to integrate all these relations [22].

$$r = f(\sum_{i,j} g([o_i; o_j; q])$$

Finally, the vector is passed through a softmax in the final layer to make the prediction to answer the question.[22]

3.2.2 Limitations of the Basic Model

Based on the basic model, we achieves a test accuracy at 55.7%. Our next step is to enhance the basic model performance to further simulate human's powerful relational reasoning ability.

In order to get a clearer view of the RNs' performance on relational tasks which contributes to human's superb communication and cognition ability, we decided to run the code in a simplified relational reasoning dataset called the Sort-of-CLEVR dataset. This dataset is introduced by the RNs proposers in their original paper [22], which is composed of paired questions and answers about images containing colorful shapes. Examples of images, questions and answers are shown as below:

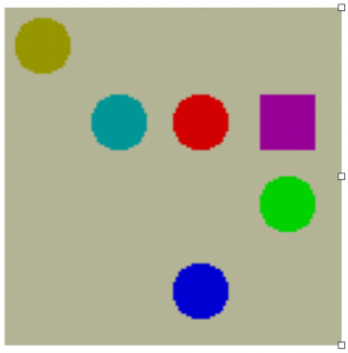


Figure 4: An example from Sort-of-CLEVR dataset. The image contains two different shapes (circle or square) with different color. An example question could be "What is the shape of the object closest to the yellow object?" and the answer would be "Circle".

The Sort-of-CLEVR dataset is interested to us as it provides with a pure test environment to reveal the model's relational reasoning ability by reducing the compounding difficulties in the following two aspects [22]:

1. The Sort-of-CLEVR dataset is a simplified version of CLEVR dataset (a 3-D visual answering dataset focusing on relational questions). The dataset contains images which are generated automatically by a generator program, where each object is randomly chosen from the defined shape types (square or circle). Each object is assigned with different color as an identifier. With an automatic image generating mechanism, the dataset removes difficulty involved in processing the pixels with actual image inputs [22].

2. The questions in Sort-of-CLEVR dataset are encoded as binary strings as a way to prevent the effect of language parsing and embedding, thus reducing the complexities with language processing [22].

To carry out our experiment, we used the code posted on GitHub by kimhc6028. The dataset in use is composed of 100000 images and 20 questions. The questions are divided into unary relational questions such as word-object recognition (e.g. What is the shape

of the green object? Circle.), binary relational questions(e.g. What is the shape of the object furthest to the blue object? Circle.) and ternary relational questions which involving relations between three objects (e.g. What is the shape of the object lies on/close the line between the purple object and the blue object? Circle.). The results for the experiment are shown as below:

Relation Types	Accuracy
Unary Relations	88%
Binary Relations	85%
Ternary Relations	55%

Table 1: Test Results on Sort-of Clevr Dataset

The results reveal that RNs perform well on tasks that fundamentally hinge on unary and binary relational reasoning. However, the result falls significantly when the number of objects required to reason across exceeds two. In other words, the original RNs are limited to perform single-step relational reasoning task as it only consider pairs of the input objects at a time as figure 5 shows. Therefore, when faced with tasks that require a chain of relational reasoning steps, the model falls short as it cannot naturally reason out the answer based on the combination or integration of all the existing relevant relations, which is also known as combinatorial entailment in RFT ideas as we mentioned.

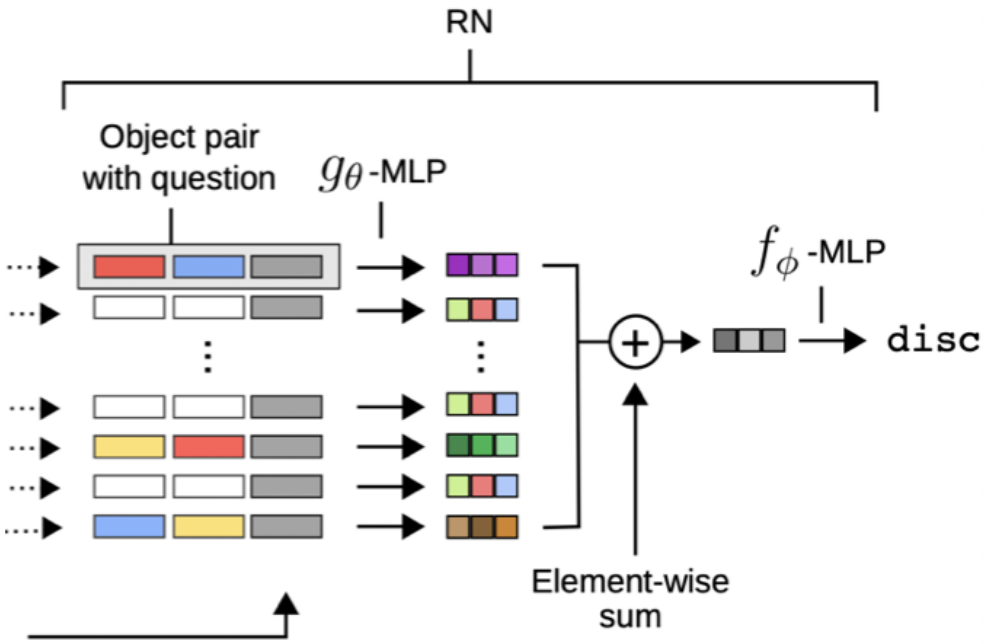


Figure 5: The RNs relational reasoning module. [22]

Another limitation that worth to mention is the scalability and applicability issue of RNs. RNs take one relation between a pair at a time among a group of objects, the number relations to be learnt is n^2 where n is the number of input objects. If n increases, the complexity increases quadratically, thus worsening the model performance as a great number of unnecessary and non-related object pairs will be operated and confuses the RNs' learning [20]. Therefore, the all-pair comparative operations presents as a major limitation of RNs in tasks which require reasoning across substantial number of information objects.[26, 5, 18]

3.3 Implementation of the Enhanced Model

3.3.1 Enhanced RNs Variations

Dealing with the scalability issue

To reduce the computational complexity from quadratic to linear while keeping the reasoning capacity of the RNs, most solutions resort to the use of attention mechanisms to filter out irrelevant relations.

Andrews et al.[5] proposed the Relationships from Entity Stream model, which as the figure 6 shows, added an entity finder RNN with a hidden state which is initially fed with the question vector to select the related entities through an attention mechanism, thus forming an entity stream. Then, the entity stream is passed to the relationship finder RNN to solve the tasks.[5]

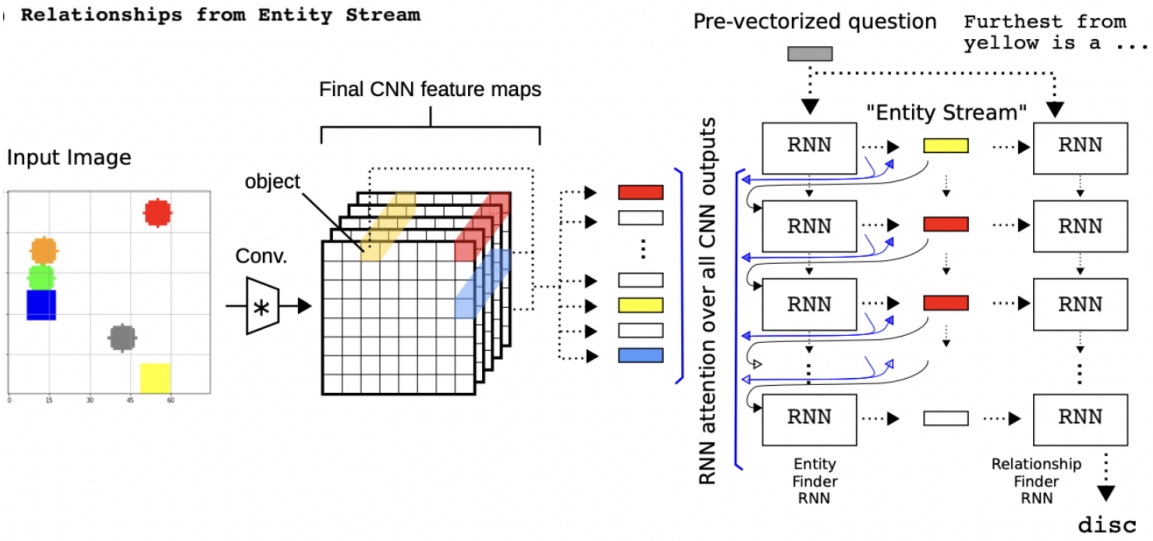


Figure 6: Relationships from Entity Stream The figure is from [5]

Other approaches such as Stochastic Relational Network [26] adapted a discriminative network to search for the objects that have highest likelihood to solve the task based on the relevance using an attention mechanism.[26]

Also, some models adapted RNs on a memory network architecture. The Relation Memory Network (RMN) [18] focuses on locating the important information with the use of MLPs to erase the information that already used. As the figure 7 shows, the model allows multiple access to renew the memory and filter our the consumed objects through an updating component in order to search for the right supporting objects among a group of entities [18].

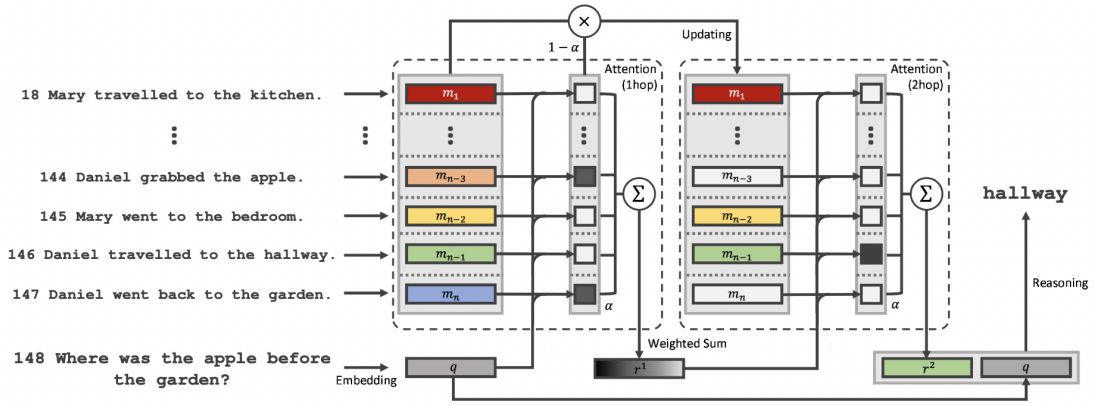


Figure 7: Relation Memory Network The figure is from [18]

Similarly, the Working Memory Network (WMN) [20] is introduced by augmenting RNs with a novel working memory storage and attentional controller module (as figure 8 shows) to filter out irrelevant inputs. The attentional controller module is designed to decide which objects should be attended. In each task, the answers that can solve the question will be seen as a condition for the attentional controller. In addition, the output of the previous hops can also serve as a condition, which makes the model to adjust focus over time in a dynamic way.[20] In this way, the attention mechanism helps to prune distractors like irrelevant objects and unnecessary relations, thereby reducing the computation costs.

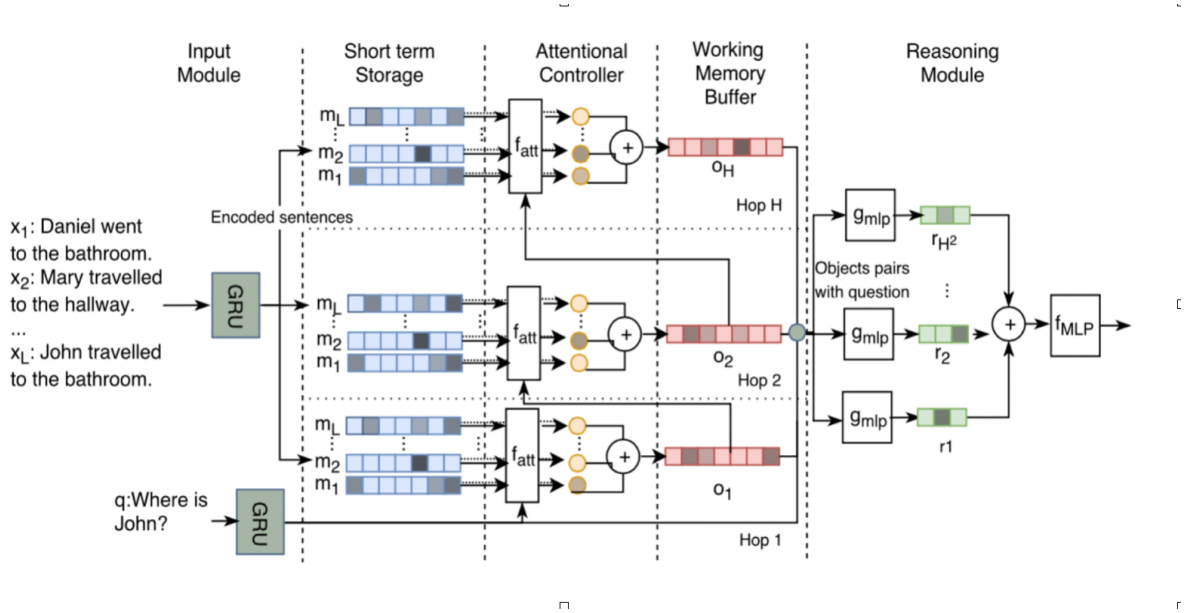


Figure 8: Working Memory Relation Networks The figure is from [20]

Dealing with the task complexity issue

In order to deal with complex relational reasoning tasks, many enhanced models seek to stack the RN layers recursively or separately, in order to achieve a transitivity of relational reasoning. The RRN (Recurrent Relational Networks) [19] considers complex relational reasoning with the use of a message passing framework. Each object input will be considered as a object with a dynamic node to record its hidden state. The initial state contains all its attributes and position of the object shown in the image. During the process, the node will be updated by its' previous state, its' input object and the state of its' neighbouring nodes through a message passing framework shown as below: [19]

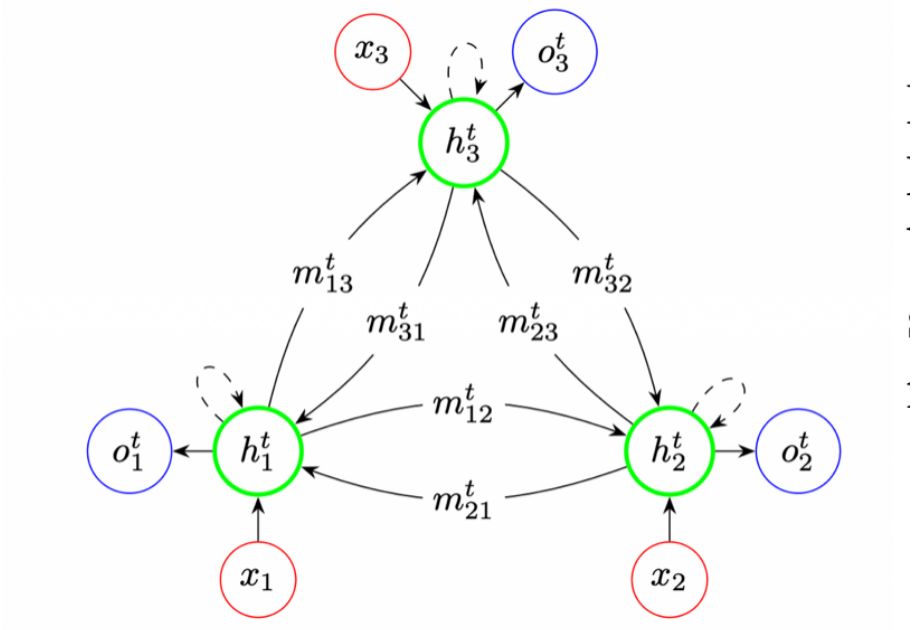


Figure 9: Recurrent Relation Networks: The network is connected by nodes. h represents the hidden state of each node which will be fed with the new inputs (indicated by the red circle x), its previous state (indicated by the dashed lines) as well as the neighboring nodes' states. The blue circles represent the outputs. The figure is from [19]

The recurrent node updating process allows the network to be repeatedly fed with messages from its' previous state, thereby allowing it to work towards a solution in an iterative way rather than starting with a blank state at every step [19]. Therefore, the network uses one recurrent cell to remember the relevant relations that have already been established, thus solving the single-layer issue of the original RNs and enabling the network to perform multi-step complex relational reasoning tasks.

However, the recurrent node architecture forces the inputs of each layer to be in the same domain, thus causing the output of all the layers to have a similar level of abstraction as the first layer's input [16]. With this weight-sharing design, the model will find it hard to learn more complex relations in deeper layers. Therefore, the Multi-layer Relation

Networks (MLRN)[16] were introduced, which simply adds more Relaiton Networks layers to the original one, which attached different weights to the outputs of each layers, thus making the complex multi-step reasoning more easily and accurately[16].

As the figure 10 shows, MLRN sum up all the relations that are related to an input object and compress them into a sum vector. In this way, the model passes one sum vector per object to the next Relation Networks layer, which in a way avoids increasing the computational complexity. Therefore, the MLRN stack the RNs layers together to form a new system, where the relations can be passed as inputs to the further layers so that more complex relations can be inferred and enable to model to be capable of dealing with tasks that require reasoning across a large number of related facts.[16]

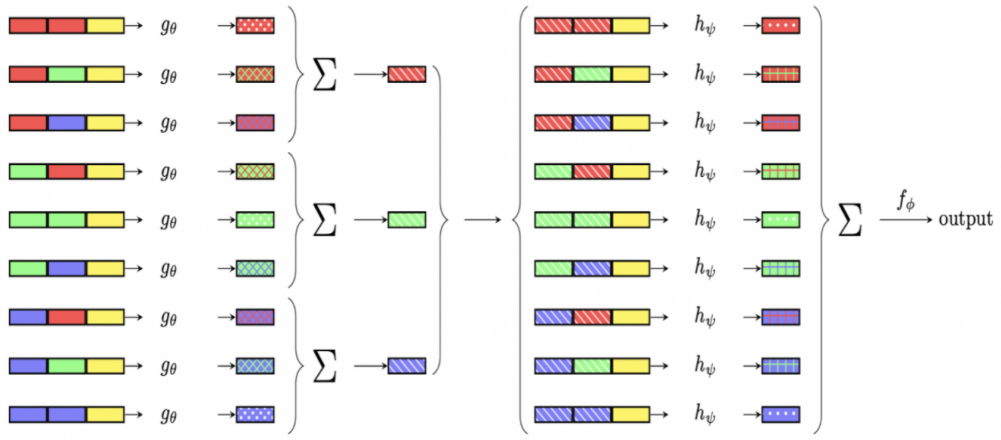


Figure 10: Multi-Layer Relation Networks: the output vectors of the first RN layer are grouped by the index i from the corresponding pairs (o_i, o_j) and are used as new inputs for another RN layer. The figure is from [16]

3.3.2 Model Design

To enhance our basic RN model and solve the problem of the task complexity issue and the scalability issue, we decided to combine the features from the MLRN[16] and WMN[20] and build our new enhanced model as the figure shows:

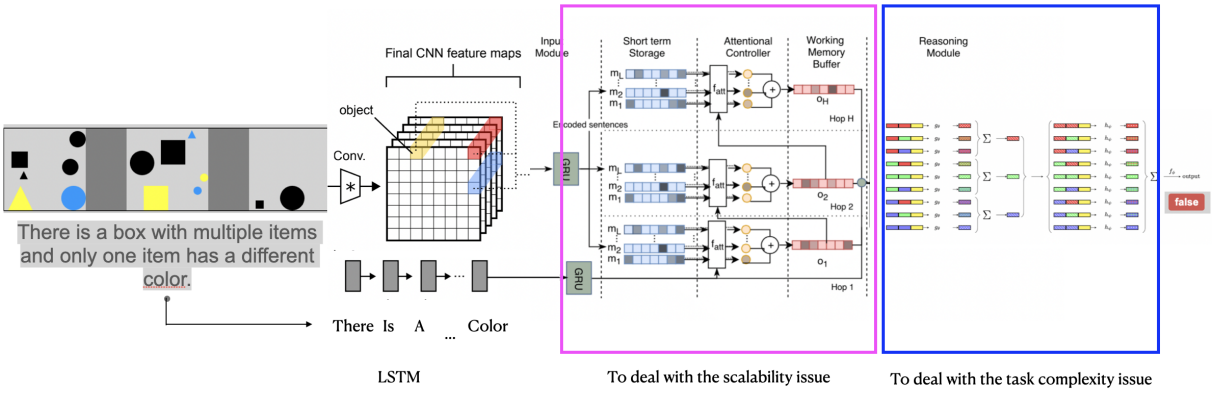


Figure 11: The Enhanced RN model applied to Natural Language for Visual Reasoning tasks. The figure is adapted from [22, 20, 16].

The enhanced model is divided into 3 major modules: the input module, the attention module and the reasoning module.

Input module The input module processes the images and textural inputs in the same way as the basic one. However, each of the output objects from the CNN is saved as a memory into a short-term storage[20].

Attention module We adapted the attention module from the WMR approach[20]. The attention module selects out the relevant objects from the short-term storage using the Multi-Head attention mechanism. During the process, the memories are projected multiple times with different projection matrices. And after each of the projection,

an attention weight will be obtained with the use of the Scaled Dot-Product attention to compute the similarity and relevance between the memory and the task question [20], and the memories in combination with the attention weight are summed up to produce a output vector. And a final output vector will be obtained which summarise all the output vectors and saved into the working memory buffer. This process will be repeated multiple times. At each time, the attention will be conditioned by the previous state through the replacement of the question vector to the previous output. And the working memory buffer will save all of these outputs from the attention mechanism before passing them to the reasoning module.[20]

Reasoning module The memories(objects) selected out from the attention module and saved into the working memory buffer will be passed as the inputs to the reasoning module. Similar to the basic model, we first used a RN layer to take the attended memories(objects) in pairs and condition them with the statement encoder. After that, we adapted the double-layer Relation Networks from Jahrens et al [16]. The architecture of the double-layer RN can be described as [16]:

$$r = f(\sum_{i,k} h(\sum_{j} g([o_i; o_j; q]); \sum_{l} g([o_k; o_l; q]); q))$$

We summarised all the established relations which reason about an object o_i and compressed them into a sum vector o_k. Then, we passed these new sum vectors as new inputs to another RN layer, thus building a double-layer Relation Network.[16]

Finally, we used the same softmax method in the final layer to answer the question.

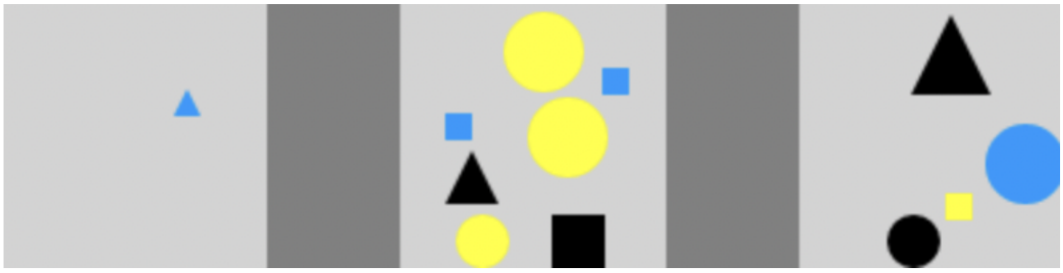
4 Results and Evaluation

4.1 the Basic Model vs. the Enhanced Model

Model Types	Test Accuracy
Basic RN model	55.7%
Enhanced RN model	60.7%

Table 2: Test results on NLVR dataset

As shown in the table, compared with the basic RN model, the enhanced RN model improves the test accuracy by about 5%, which proves that the enhanced RN model performs better in complex visual reasoning tasks. For instance, both models can easily learn to judge the sentences focusing on unary (see figure 12) or binary relations (see figure 13).



There are two blue blocks. Answer: True

Figure 12: Unary Relation Example from NLVR Dataset

There is a blue block above a yellow block. Answer: True

Figure 13: Bnary Relation Example from NLVR Dataset

However, the enhanced model improves the accuracy of tasks that require a combinatorial reasoning of multiple relations or require reasoning across more than 2 supporting facts, shown as the figure below:

There is exactly one tower with a blue block over a yellow block. Answer: True

There are two towers with a blue block over a blue block. Answer: True

Figure 14: Ternary Relation Example from NLVR Dataset

The reason of such improvements lies in the combination of attention mechanism [20]

and double-layer RN design [16]. With the attention mechanism, the model is able to filter out the irrelevant memory and search for the relevant parts to produce the output, thereby locating the supporting facts more precisely and effectively [20]. With the double-layer RN design, the model learns to compress the relevant established relations into vectors and combine them to reason out the output, thus performing multi-step relational reasoning tasks [16]. With these new features combined, the model improves the accuracy on these complex reasoning tasks.

In addition, compared with the basic model, the enhanced model reduces the computation complexity significantly. As we mentioned before, the original RNs take every pair of objects at a time, which scales quadratically if the number of objects increase. With the adoption of WMN [20], the running time is dependent of the size of the working memory buffer, which is much smaller than the number of objects as it will filter out the irrelevant objects with the attention weight [20].

4.2 Comparison with Other Models

According to the leaderboard from NLVR official webiste [1], most of the leading methods approach the visual reasoning with the use of structured representations of the image inputs [10, 13, 11]. However, the RNs model specializes with its ability to process the raw visual inputs directly with the use of a CNN module, with which the model is allowed to define what constitute an object flexibly during the learning process [22]. In this way, the model can recognize suitable objects from different settings and learn any type of relations between these objects. Therefore, our approach differs as it is able to process the visual inputs through pixels directly instead of using the structured representations inputs.

The table below shows the leaderboard of the models which also use the raw image inputs directly to process the tasks [1].

Rank	Model	Test Accuracy
1	CNN-BiATT	69.7%
2	N2NMN	69.1%
3	Neural Module Networks	66.1%
4	FiLM	62.2%
5	MAC-Network	57.6%
6	Majority Class	56.2%

Table 3: NLVR Leaderboard (using images)[1]

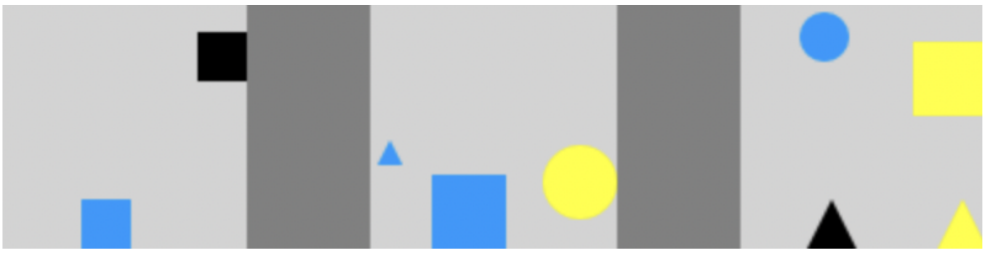
We decided to compare our RNs approach with other models having similar results shown in the table such as Neural Module Networks [3], FiLM [21] and MAC-Network [15]. Different from the Neural Module Networks [3] which uses a natural language parser to dynamically restructure and decompose the statements into logical and semantic expressions before passing them into the neural networks, our approach takes the raw human-written sentences as inputs and trains the model end-to-end without the use of natural language processing tools to prepare the raw textual inputs into structured chunks [20]. In this way, the model is able to process and interpret the statements on its own during the training process.

FiLM [21] and MAC-Network [15] both approach the complex underlying reasoning structure of the tasks with the use of conditioning methods. The statements are decomposed into a series of conditions or operations, thus retrieve or locating the information from the visual input in a selective and effective manner [21, 15]. However, our approach features as we train the model to capture the interactions between the visual objects with attributes. Therefore, the model is able to learn the relevant relations from the images before answering the question, which is more like a close-book exam instead of searching for the answer after the question provided. In other words, our model is trained to learn the relations through a reasoning module, while the other two models learn to find the answer during the process. From the perspective of RFT, our model approach the task through the learning process which simulates our human brains as

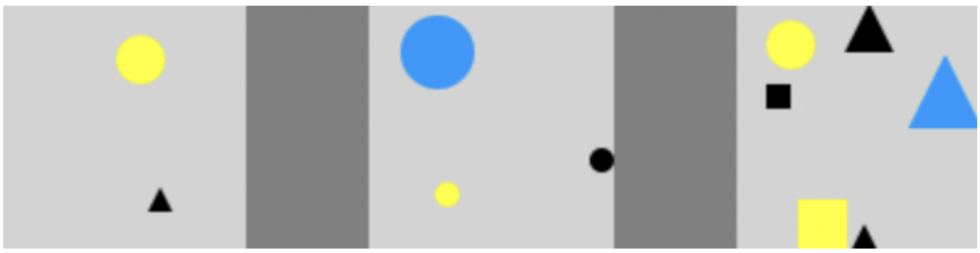
humans learn to produce or understand novel sentences through creating or identifying associations and links between the words in statements and the visual objects as well as the physical or abstract relations (i.e. spatial, quantitative relations) between these visual objects. Therefore, our approach specializes as it simulates human's language learning process with the ability to learn the underlying comparisons and relations before answering the question.

4.3 Limitations

Given the fact that our approach processes the textual input by itself without the use of NLP (Natural Language Processing) tools to transfer them into structured representations, the model is limited to process high-level complex phrases appearing in the statements. However, as the statements are written by different people, it is inevitable that different writers might use different by complex English phrases to describe the same meaning. For example, in figure 15, the statements contains phrases like "touching the wall" and "touching the edge". These phrases impose challenges for our end-to-end model to understand and connect them with correct visual information.



There are exactly three blue objects not touching any edge. Answer: False



There is a small black triangle touching the wall. Answer: True

Figure 15: Incorrectly Answered Examples from NLVR Dataset

It is because that understanding these expressions requires the model to process a combination of a group of certain words. However, people will change these word combinations but to refer to the same meaning, which bring difficulty for the model to process. Another reason is these high-level complex phrases didn't appear frequently and sufficiently in the dataset which will cause a lack of training data to recognize such patterns.

5 Conclusion

In this paper, we approached the visual reasoning tasks through incorporating relational reasoning ability into the neural network. According to RFT [12], humans' unique and superb cognition and language ability is founded in our powerful relating ability, which

enables us to identify, create or derive bidirectional associations between stimuli like words, events or ideas. Inspired by RFT ideas, we adapted the Relation Networks [22] as our base reasoning module as it simulates human brain's relating ability through considering the potential relations between all object pairs, recognising the objects during the learning process, inferring the relations from the training and being able to capture both abstract and physical relations between objects. After running the adapted RNs model in the NLVR dataset, we achieves the accuracy of 55.7% in the test data, which shows a relatively high performance according to the leaderboard from the NLVR website [1]. However, the basic model suffers from the scalability problem as the computation cost will increase quadratically if the number of objects increases. In addition, the basic RNs model falls short dealing with tasks that requires the reasoning across more than two fact inputs as it is designed to consider a pair of objects at a time. In order to cope with the scalability issue and the task complexity issue, we combined the attention mechanism from the Working Memory Networks [19] to filter out the irrelevant inputs and the multi-layer design from the Multi-layer Relation Networks [16] to combine the established relations together in order to reason out the result. These enhancement features improves our model by 5% accuracy in the test data, which achieves 60.7% accuracy. Compared with other models with similar results, our approach features as we process the textual and visual inputs directly without transferring them into structured representations before passing into the model. Furthermore, our approach specialised as we enable the model to focus on learning the relations between objects during the training process and have a command of the relevant relational knowledge between objects in the image, which means the model is trained to answer the question as it knows the answers beforehand instead of learning to find the answer after the question presented. However, our model is also limited to process some high-level complex phrases written by humans due to the arbitrariness of word combinations and the lack of training data.

Overall, the project demonstrates the feasibility of applying RFT ideas into machine's cognition building or language learning process. Future works can be done to implement the model into more authentic language learning environment and test the model's

recognition and language performance based on relational reasoning ability in more realistic context.

Find our source code at https://github.com/Lizhao-Liu/Relational-Reasoning-Project.

References

[1] Natural language for visual reasoning. https://lil.nlp.cornell.edu/nlvr/. Accessed: 2021-07-30.

[2] What is relational frame theory? a psychologist explains. https://positivepsychology.com/relational-frame-theory/. Accessed: 2021-07-30.

[3] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Deep compositional question answering with neural module networks. corr abs/1511.02799 (2015). *arXiv preprint arXiv:1511.02799*, 2015.

[4] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48, 2016.

[5] Martin Andrews and Sam Witteveen. Relationships from entity stream. *arXiv preprint arXiv:1909.03315*, 2019.

[6] Steven C Hayes Dermot Barnes-Holmes and Bryan Roche. Relational frame theory: A post-skinnerian account of human language and cognition. 2001.

[7] Peter W Battaglia, Razvan Pascanu, Matthew Lai, Danilo Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. *arXiv preprint arXiv:1612.00222*, 2016.

[8] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[9] Veronica Cullinan and Agata Vitale. The contribution of relational frame theory to the development of interventions for impairments of language and cognition. 2009.

[10] Pradeep Dasigi, Matt Gardner, Shikhar Murty, Luke Zettlemoyer, and Eduard Hovy. Iterative search for weakly supervised semantic parsing. In *Proceedings*

of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2669–2680, 2019.

[11] Omer Goldman, Veronica Latcinnik, Udi Naveh, Amir Globerson, and Jonathan Berant. Weakly-supervised semantic parsing with abstract examples. *arXiv preprint arXiv:1711.05240*, 2017.

[12] Amy C Gross and Eric J Fox. Relational frame theory: An overview of the controversy. *The Analysis of verbal behavior*, 25(1):87–98, 2009.

[13] Nitish Gupta, Sameer Singh, and Matt Gardner. Enforcing consistency in weakly supervised semantic parsing. *arXiv preprint arXiv:2107.05833*, 2021.

[14] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

[15] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018.

[16] Marius Jahrens and Thomas Martinetz. Multi-layer relation networks. *arXiv preprint arXiv:1811.01838*, 2018.

[17] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[18] Jihyung Moon, Hyochang Yang, and Sungzoon Cho. Finding remo (related memory object): A simple neural architecture for text based reasoning. *arXiv preprint arXiv:1801.08459*, 2018.

[19] Rasmus Berg Palm, Ulrich Paquet, and Ole Winther. Recurrent relational networks. *arXiv preprint arXiv:1711.08028*, 2017.

[20] Juan Pavez, Héctor Allende, and Héctor Allende-Cid. Working memory networks: Augmenting memory networks with a relational reasoning module. In *Proceed-*

ings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2018.

[21] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[22] Adam Santoro, David Raposo, David GT Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *arXiv preprint arXiv:1706.01427*, 2017.

[23] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[24] Ian Stewart, John McElwee, and Siri Ming. Language generativity, response generalization, and derived relational responding. *The Analysis of verbal behavior*, 29(1):137–155, 2013.

[25] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, 2017.

[26] Kang Min Yoo, Hyun Soo Jo, Hanbit Lee, Jeeseung Han, and Sang-goo Lee. Stochastic relational network. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 788–792. IEEE, 2019.