INTERNATIONAL JOURNAL OF COMPUTERS
DOI: 10.46300/9108.2020.14.9 Volume 14, 2020

Enlightening and Predicting the Correlation Around Deep Neural Nets and Cognitive Perceptions

Chandra Bhim Bhan Singh
Kurukshetra, Haryana, India

Abstract: **Recently, psychologist has experienced drastic development using statistical methods to analyze the interactions of humans. The intention of past decades of psychological studies is to model how individuals learn elements and types. The scientific validation of such studies is often based on straightforward illustrations of artificial stimuli. Recently, in activities such as recognizing items in natural pictures, strong neural networks have reached or exceeded human precision. In this paper, we present Relevance Networks (RNs) as a basic plug-and-play application with Covolutionary Neural Network (CNN) to address issues that are essentially related to reasoning. Thus our proposed network performs visual answering the questions, super-human performance and text based answering. All of these have been accomplished by complex reasoning on diverse physical systems. Thus, by simply increasing convolutions, (Long Short Term Memory) LSTMs, and (Multi-Layer Perceptron) MLPs with RNs, we can remove the computational burden from network components that are unsuitable for handling relational reasoning, reduce the overall complexity of the network, and gain a general ability to reason about the relationships between entities and their properties.**

Keywords: **Cognitive science, Artificial Intelligence, Resemblance, Object Classification, Neural Networks, Convex Optimizatio**n

I. INTRODUCTION:

Cognitive science is defined as mental and brain scientific research, such as psychiatry, mental philosophy, neurology, anthropology, sociology, informatics, and robotics [1]. This science examines the nature of mental operations that make these actions possible, such as thinking, classification and procedures [2]. More specifically, vision, thinking and reasoning, memory, attention, learning, and language-related subjects are the primary objectives of this study [3] [4]. Cognitive psychology is regarded as a branch of cognitive science that examines the mind's internal procedures like critical thinking, consciousness, awareness, cognition, language, and problem-solving [5]. Recently, psychological research has studied different kinds of cognitive psychology scopes including science techniques, qualitative perception, coordination of quantitative statistics, and the descriptive hypothesis [6] [7]. The ability to reason about the relationships between entities and their properties is central to smart behavior in general (Figure 1) [8]. Consider a child proposing a race between the two park trees that are farthest apart: it is necessary to infer the pair distances between each tree in the park and compare it to know where to run. Or, consider a reader gathering evidence to predict the culprit in a murder-mystery novel: to build a plausible narrative and solve the mystery, each clue must be considered in its broader context [9].

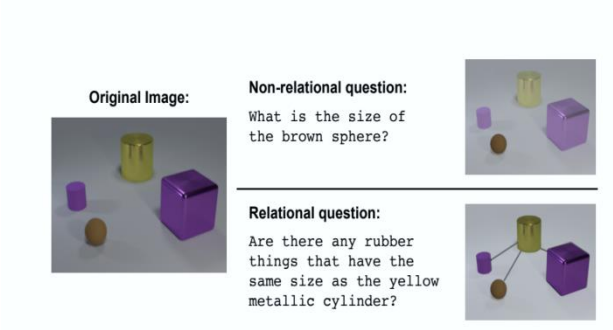

Figure 1: illustration of relational reasoning

Artificial intelligence symbolic methods are fundamentally related. Practitioners describe the relationships between symbols using the language of logic and mathematics, and then use a multitude of effective methods to explain these relationships, including deduction, arithmetic, and algebra [10].

But symbolic approaches suffer from the problem of symbol grounding and are not robust to small variations in tasks and inputs [11]. Certain methods, such as those based on statistical learning, construct raw data representations and often generalize through complex and noisy conditions. Nevertheless, a number of such methods, such as deep learning, also struggle in data-poor issues where sparse yet complex relationships characterize the underlying structure [12]. Thus, existing methods often struggle with structured and combinatorial issues, so there is a great need to introduce a novel solution. Our findings support these statements and further show that apparently simple relational inferences are remarkably difficult for strong neural network architectures such as convolutionary neural networks (CNNs) and multi-layer perceptron's (MLPs) [13]. Here we are exploring "Relevance Networks" (RN) as a general solution in neural networks for relational reasoning. RNs are architectures that explicitly focus on relational reasoning in their calculations [14]. Although several other models have been proposed that endorse relationship-centered computation, such as Graph Neural Networks [15], Gated Graph Sequence Neural Networks [16], and Interaction Networks [17], RNs are simpler, more exclusively focused on general relationship reasoning, and more easily integrated into broader architectures.

The rest of paper is organized as follows. In Section 2, survey of previous literature is addressed. In section 3, our proposed frame work is presented and discussed. Section 4, gives the result and evaluate the performance of our proposed method under various scenarios. Finally, we conclude the paper in section 5.

II. LITERATURE SURVEY:

Henaff et al. [18] The Recurrent Entity Network (EntNet) introduces a new paradigm. It is fitted with a complex long-term memory that enables a representation of the state of the world to be preserved and modified as it receives new data. It can reason on - the-fly for language comprehension tasks as it reads text, not just when it is necessary to answer a question or response as is the case with a memory network. Like a Neural Turing Machine or Differentiable Neural Computer, it maintains a fixed size memory and is able to learn how to read and write operations based on location and content.

Johnson et al. [19] we need diagnostic tests to assess our progress and identify vulnerabilities when developing artificial intelligence systems that can reason and answer questions about visual data. We often conflate several sources of error, making it difficult to find flaws in the model. We present a dataset of diagnosis that measures a range of abilities in visual reasoning. It includes minimal biases and has extensive annotations that explain the type of reasoning needed by each query.

Hariharan et al. [20] Proposes a visual reasoning model consisting of a program generator that provides an explicit representation of the reasoning process to be performed and an execution engine that performs the resulting program to provide a response. Neural networks implement both the program generator and the execution engine and are trained using a back propagation and REINFORCE combination.

Kafle et al. [21] analyze existing VQA algorithms using the new Task Driven Image Understanding Challenge (TDIUC) dataset, which has over 1.6 million questions in 12 different categories. We often add meaningless questions for a given image in order to force a VQA framework to think about the quality of the image. We are proposing new assessment systems that make it easier to research the strengths and weaknesses of algorithms to account for over-represented problem forms. We evaluate both baseline and state-of - the-art VQA models output, including multi-modal compact bilinear pooling (MCB), neural node networks, and recurrent response units.

Lake et al. [22] described cognitive science advancement means that genuinely human-like learning and reasoning machines need to go beyond current trends in engineering in both what they know and how they learn it. Specifically, we argue that these machines can (a) construct causal world models that support interpretation and comprehension rather than simply solving pattern recognition problems; (b) ground learning in intuitive physics and psychology theories to help and expand the information learned; and (c) make use of compositionality and learning-to-learn to rapidly acquire and generalize knowledge.

Malinowski et al. [23] propose a Deep Learning approach to answering the visual query challenge, where machines respond to real-world picture questions. By incorporating the latest developments in representation of images and the processing of natural languages, we propose Ask Your Neurons, a scalable, jointly trained end-to-end solution for this problem. Unlike previous efforts, we face a multi-modal problem where the language output (answer) depends on visual and natural language inputs (image and question).

In [18] does not get greater accuracy, [19] system have more complexity [20] describe important categorize information but fail to finite representations [21] classification have complication in structure [22] attains more memory intricacy [23] classification with less objective classes. Thus to overcome above mentioned challenges, there is a great need to develop a novel methodology.

III. RELATIONAL REASONING WITH RELEVANT NETWORK:

An RN is a node of the neural network with a structure based on logical reasoning. The design philosophy behind RNs is to limit a neural network's functional form in order to capture the core common properties of relational reasoning. In other words, the ability to measure relationships is built into the RN architecture without understanding, just as the ability to reason spatial, translation invariant properties is built into CNNs, and the ability to reason sequential dependencies is built into recurrent neural networks.

The RN is a composite function in its simplest form:

$$RN(C) = f_{\phi}\left(\sum_{i,j} g_{\theta}(c_i, c_j)\right) \quad (1)$$

Where the input is a set of "objects" C={ c1,c2,....., cn}, ci € Rm is an ith object, and $f_{\phi} and g_{\theta}$ are both function with parameters ϕ and θ respectively. We are MLPs for our purposes, and the parameters are synaptic weights that can be taught, making RNs end-to-end differentiable. We call the production of g_{θ} a "relationship;" therefore, the function of g_{θ} is to infer how two objects are connected, or whether they are related at all.

RNs have three notable strengths: they learn to infer relationships, they are data-efficient, and they work in an invariant order format on a collection of objects–a highly common and flexible input format. The inputs are given into training phase to predict the interactions among the objects. The Relevant Network's training process is detailed in the next section.

A. RNs train for predict interactions:

The functional form in Equation 1 specifies that the possible relationships between all object pairs should be interpreted by an RN. This means that an RN is not inherently private to which object relations actually exist, or to any particular relationship's actual meaning. RNs must therefore learn to infer the nature and consequences of relationships with objects.

In graph theory parlance, the input can be seen as a complete and directed graph whose nodes are objects and whose edges represent the pairs of objects whose relationships should be taken into account. Although we concentrate throughout this paper on this "all-to-all" variant of the RN, this RN description can be modified to include only a few pairs of objects. Similar to interaction networks associated with RNs, RNs will input a list of only those

pairs that should be considered if this information is available. This knowledge might be evident in the input data, or some upstream process may extract it.

B. RNs provide effective in data:

To measure each relationship, RNs use a single g_θ function. This can be viewed as a single function operating on a batch of pairs of objects, where each batch member is a different pair of objects from the same collection of objects. This mode of operation promotes greater generalization of device relationships as g_θ is encouraged not to over-adjust to any single object pair's features. Consider how the same function would learn from an MLP. An MLP will receive all objects simultaneously as their input from the defined object. In order to account for all possible object pairings, it must learn and embed n2 (where n is the number of objects) similar functions within its weight parameters. As the number of objects increases, this soon becomes intractable. The cost of learning a relationship function n2 times using a single feed forward pass per sample, as in an MLP, is therefore replaced by the cost of n2 feed forward passes per object set (i.e. for each possible object pair in the set) and learning a relationship function only once, as in an RN.

The summation in Equation 1 guarantees that the RN is invariant in its input to the order of things, following the property that sets invariant in order. Because we used summation, it is possible to use other commutative operators instead, such as max and average pooling. Therefore, a neural network approach has been introduced to that is incorporated with the RN. The structural explanation is given in the following sections.

C. Visual Architecture of neural network:

RNs operate on objects in their simplest form, and therefore do not operate directly on images or natural language. A key contribution of this work is to show the flexibility with which relatively unstructured inputs, such as CNN or LSTM embedding, can be considered as a set of objects for an RN. As described below, in factorizing the input of the RN into a set of objects, we require minimal oversight.

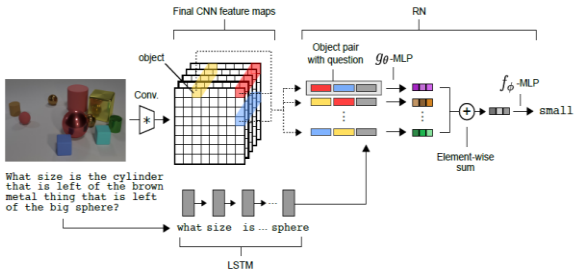

Figure 2: Visual Architecture of neural network

In figure 2 questions are processed with an LSTM to generate an embedding query, and images are processed with a CNN to create a set of RN objects. Objects (three examples outlined in yellow, red, and blue) are constructed from the transformed image using feature-map vectors. The RN considers relationships across all pairs of objects, depending on the embedding of the query, and incorporates all these relationships to answer the question.

We used a CNN to interpret a collection of objects with pixel inputs. The CNN took 128x 128 images and converted them into k feature maps of size d x d through four convolution layers, where k is the number of kernels in the final convolution layer. We remained agnostic about what specific features of the image should be an object. Thus, each of the d2 k-dimensional cells in the d x d feature maps was tagged with a coordinate (from the range (-1, 1) for each of the x-and y-coordinates) 3 indicating their relative spatial position, and treated as an object for the RN. Thus the process is clearly explained in figure 2. This means that an object could include the context, a particular physical object, a texture, physical object conjunctions, etc., which in the learning process allows the model great flexibility.

An object-object relationship's life and purpose should be question-dependent. For example, if a question asks about a large sphere, it is likely that the relationship between small cubes is meaningless.

We changed the RN architecture so that g_θ could make its processing conditional on the question:

$$a = f_\phi\left(\Sigma_{i,j}\, g\theta\left(c_i, c_j, q\right)\right) \quad (2)$$

We used the final state of an LSTM that interpreted query terms to get the question embedded q. Different integers were assigned to the query terms, which were then used to index an apprenticed lookup table that supported LSTM embedding. According to the English-encoded query syntax, the LSTM received a single word embedding as input at each time-step.

We can provide state descriptions directly in the RN, as state descriptions are pre-factored representations of objects. Question processing will continue as before: questions move through an LSTM using a learnable lookup embedded for individual words, and each object-pair is concatenated with the final state of the LSTM. Then the natural language system based on text QA is explained in detail below.

D. Natural Language system:

The natural language inputs have to be transformed into a set of objects for the text based QA of tasks. This is a distinctly different requirement from visual QA, where objects in converted feature maps were identified as spatially separate regions. So, we first took in the help package the 20 sentences that were immediately before the issue of the investigation. Then we marked these sentences with labels showing their relative position in the support set and treated each word-by-word sentence with an LSTM (with the same LSTM acting independently on each sentence). We note that this configuration invokes limited prior knowledge by delineating artifacts as sentences, while previous models processed all word tokens sequentially from all help sentences.

How much benefit this prior knowledge offers is uncertain, because time punctuation often clearly delineates sentences for token-by-token processing models. The sentence-processing-LSTM's final state is called an entity. Similar to visual QA, a separate LSTM generated an

embedding query that appeared as an input to the RN for each pair of objects. We note that this configuration invokes limited prior knowledge by delineating artifacts as sentences, while previous models processed all word tokens sequentially from all help sentences.

How much benefit this prior knowledge offers is uncertain, because time punctuation often clearly delineates sentences for token-by-token processing models. Sentence-processing-LSTM's final state is called an entity. Similar to visual QA, a separate LSTM generated an embedding query that appeared as an input to the RN for each pair of objects, the elaborate configuration of visual mode is follows.

E. Visual Model configuration:

For the pixel task, we used: 4 convolution layers with 24 kernels each, ReLU non-linearity's and batch normalization; 128 LSTM units for query processing; 32 word-look-up embedding units; 4-layer MLP consisting of 256 units per layer with ReLU non-linearity's for g_θ; and 3-layer MLP consisting of 256, 256 (with 50 percent dropout) and 29 units with ReLU non-linearity's for f_ϕ. The final layer was a linear layer over the solution vocabulary generating logics for a soft max. The soft max performance was optimized using the Adam optimizer with a learning rate of 2:5 with a cross-entropy loss function. Compared to the visual QA architectures used, we would like to emphasize the simplicity of our overall model architecture using ResNet or VGG embedding, sometimes with fine tuning, very large LSTMs for language encoding, and further processing modules, such as stacked or iterative attention, or wide fully connected layers (over 4000 units, often).

Therefore from the above techniques with neural models for relational reasoning performance based on visual based QA, text based QA and dynamic physical based system is analyzed and experiment is conducted between efficient datasets. Thus the performance analysis is described in next section.

IV. Result and Discussion:

We applied RN-enhanced networks to a range of tasks that rely on relational reasoning. We selected tasks from several different domains, including visual QA, text-based QA, and diverse physical structures, to show the flexibility of these networks.

A. Dataset description:

A model needs to learn how to answer questions about an image in visual QA. This is a daunting problem domain as it includes comprehension of the high-level scene. Architectures will perform complex conceptual reasoning–spatial and otherwise–about the characteristics of visual inputs, language inputs, and their conjunction. Nevertheless, in the absence of clearly defined word vocabulary, the majority of visual QA datasets involve logic, and perhaps more perniciously, vast and complicated world knowledge that is not accessible in the training data.

They also contain ambiguities and exhibit strong linguistic biases that enable a model to learn responsive strategies that exploit those biases without rationalizing visual input.

The CLEVR visual QA dataset was developed to control these issues and distill the core challenges of visual QA. CLEVR contains representations of objects made in 3D, such as spheres and cylinders. A number of questions that fall into different categories are associated with each image. For example, questions from the query attribute may ask "What is the sphere color?" Is the cube the same material as the cylinder, while comparing attribute questions may be asked?" An important feature of CLEVR for our purposes is that many problems are directly linked in nature. Remarkably, efficient QA frameworks can't solve CLEVR, probably because they can't handle the task's core relevant aspects.

1. Visual based QA:

In order to explore our hypothesis that the RN architecture is better suited to general relational reasoning compared to more traditional neural architectures, we designed a CLEVR-like dataset called "Sort-of-CLEVR." This dataset divides related and non-related issues.

Sort-of-CLEVR consists of 2D colored shapes images along with image questions and answers. Each image has a total of 6 objects, where each object is a form (square or circle) chosen randomly. To classify each object unambiguously, we used 6 colors (red, blue, green, orange, and yellow, gray).

Questions are hard-coded as fixed-length binary strings to reduce the difficulty involved in the processing of natural language question-words and thus eliminate any confusing problem with language parsing. 10 relational questions and 10 non-relational questions were created for each image. Examples of relational questions are: "What is the form of the object as far as the gray object is concerned? "And" How many objects have the form of a green object?". Examples of non-relational issues are: "What is the gray object's shape?" And is the blue entity at the top of the scene or at the bottom?".

2. Text based QA:

BAbI is a pure QA dataset based on text. There are 20 tasks, each of which refers to a different type of reasoning, such as inference, induction, or counting. Each problem is connected to a set of facts that support it. The "Sandra picked up the football" evidence, for example, and "Sandra went to the office" support the question "Where is the football?"(Answer:" office). A model succeeds in a mission if its output reaches 95%. Several neural networks that have improved their memory have recorded impressive results on bAbI.

3. Dynamic and physical based QA:

Using the MuJoCo physics engine, we built a data set of simulated physical mass-spring systems. There were 10 colored balls moving on a table-top surface in each sequence. Some of the balls were moving freely, free to interfere with other balls and walls of the barrier. Certain ball pairs randomly selected were bound by invisible springs or a rigid restriction. Due to the force exerted by the links, these connections stopped the balls from moving independently. Input data consisted of state definition matrices in which each ball was represented as a row in a matrix with characteristics reflecting each object's RGB

color values and their spatial coordinates (x and y) over 16 sequential time step.

The implementation of random ball-to-ball connections created an emerging physical system with a variable number of connected ball "systems" (where "systems" refers to connected ball graphs as nodes and ball-to-edge connections). We specified two separate tasks: 1) infer the presence or absence of ball connections only when observing their color and coordinating positions across multiple sequential frames, and 2) count the number of systems on the table top, again when observing the color and coordinating position of each ball across multiple sequential frames. Both of these tasks include thinking about the balls ' relative positions and velocities to decide whether they are moving independently or whether their movement relies on other balls ' movement through invisible connections. For instance, if the distance between two balls remains identical across frames, a relation between them can be inferred. The first task makes such inferences clear, while the second task allows tacit, much more complicated, reasoning to occur.

B. Performance Evaluation:

Our model achieved state-of - the-art performance is followed by, exceeding by 27 percent the best model trained only on the pixel images and questions when publishing the data set, and exceeding human performance in the task.

These results–especially those obtained in the attribute comparison and count categories–are a testament to our model's ability to make relational reasoning. In fact, the most struggles between state-of - the-art models is in these categories. In addition, the relative simplicity of the network components used in our model indicates that the CLEVR task's difficulty lies in its associated reasoning demands, not in the language or visual processing.

1. Length Distribution:

The main statistics for visual based QA is the majority of questions are unique and few questions from the validation and test sets appear in the training set. In our proposed work used the questions are much longer compared to other existing techniques.

Table 1: Length Distribution Performance For Clevr

Si. No	Question fraction	Words per question (%)
1	2.5	19.5
2	10	15.9
3	15	9.86
4	18	6.45
5	20	6.99
6	25	5.423
7	30	2.489
8	35	1.579
9	40	0.96

In table 1 described the length distribution performance of CLEVR dataset. Here, question fractions are 2.5, 10, 15, 18, 20, 25, 30, 35, and 40 respectively and for corresponded words per question is 19.5, 15.9, 9.86, 6.45, 6.99, 5.423, 2.489, 1.579, and 0.96%. And the graphical illustration of table 1 is in figure 3.

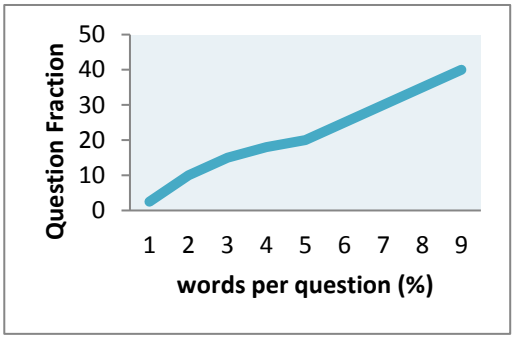

Figure 3: Length Distribution performance

2. Accuracy:

Most visual problems include computing and comparing more than one relationship; for instance, consider the question: "There's a big thing on the right side of the big rubber cylinder behind the big cylinder on the right side of the small yellow rubber thing; what's its shape? “Who has three spatial relationships" (the right side," "the right side," "the right side). Our model achieves output at 98 percent on such queries, suggesting that the model can handle complex relational reasoning.

Table 2: Performance Analysis Of Our Proposed Technique

Si.No	Proposed performances	Measurements (%)
1.	Accuracy	96.8
2.	Data Count	92.2
3.	Exist data’s	98.6
4.	Compare numbers	94.7
5.	Query Attribute	98.8
6.	Compare Attribute	98.2

In table 2 represents the performance analysis based on accuracy, which described the efficient concert of our proposed technique during relational reasoning exploration. Here, our system performs the data count is 92.2%, exist data’s are 98.6%, compared numbers are 94.7%, query attribute is 98.8%, compare attribute is 98.2% and the overall accuracy is 96.8% respectively, thus graphical illustration is represented in figure 4.

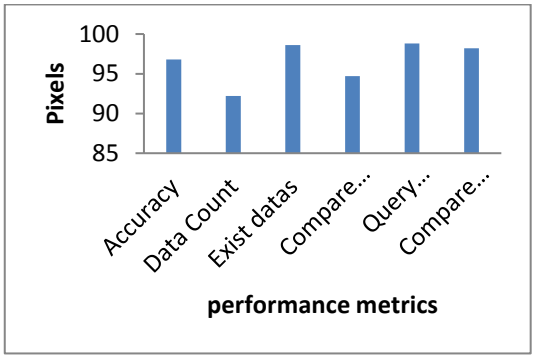

Figure 4: Overall accuracy of proposed technique

Table 3: Numbers Comparing Analysis

Models	Overall	Data Count	Exist data's	More than	Less than	Equal
Human	0.91	0.81	0.91	0.78	0.81	0.78
Q-type base line	0.45	0.31	0.51	0.53	0.55	0.57
LSTM	0.49	0.36	0.59	0.68	0.62	0.71
CNN+LSTM	0.52	0.35	0.63	0.61	0.69	0.66
CNN+LSTM +SA	0.65	0.51	0.69	0.71	0.73	0.51
CNN+LSTM +RN	0.99	0.98	0.99	0.98	0.96	0.91

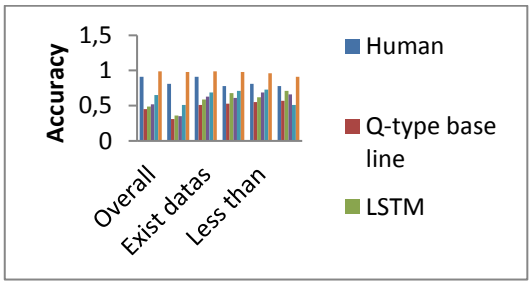

Figure 5: graphical illustration for comparing the numbers

Table 4: Performance For Compare Attributes And Query Attribute

Models	Query size	Query shape	Query material	Query colour	Compare size	Compare shape	Compare material	Compare color
Human	0.88	0.88	0.87	0.91	0.85	0.98	0.91	0.99
Q-type base line	0.45	0.28	0.51	0.11	0.55	0.53	0.57	0.55
LSTM	0.46	0.28	0.53	0.13	0.57	0.55	0.59	0.55
CNN+LSTM	0.51	0.49	0.55	0.25	0.61	0.56	0.59	0.58
CNN+LSTM+SA	0.79	0.78	0.79	0.75	0.59	0.59	0.59	0.58
CNN+LSTM+RN	0.99	0.97	0.98	0.99	0.97	0.98	0.99	0.97

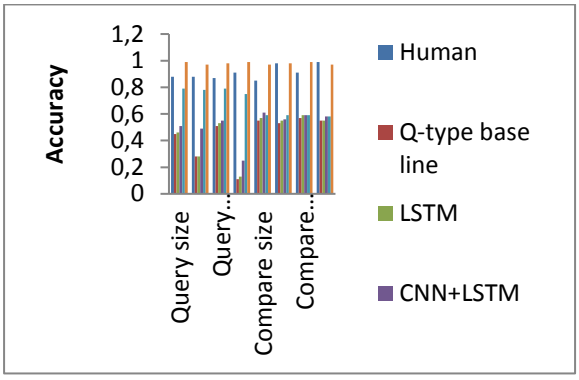

Figure 6: graphical illustration of performance for compare attributes and query attribute

C. Comparison Analysis:

A more recent study reported 98 percent overall performances on CLEVR, but used additional monitoring signals on the functional programs used to generate the CLEVR questions. It is not possible for us to compare this directly with our work as we do not use these additional signals of supervision. Nevertheless, our methodology outperforms a version of their model that has not been conditioned with these extra signals, and even a version of their model equipped with 9 K ground-truth programs. RNs can therefore produce very competitive and even super-human results under much weaker and more normal assumptions, even in circumstances where usable programs are not available.

Table 5: Comparison Of Accuracy

Si.No	Models	ACCURACY
1.	Human	93.1
2.	Q-type base line	42.5
3.	LSTM	48.5
4.	CNN+LSTM	53.7
5.	CNN+LSTM+SA	69.6
6.	CNN+LSTM+ improved SA	77.1
7.	proposed technique	98.5

In table 5 described the performance of proposed technique accuracy with existing, thus it clearly described, our technique attains very high accuracy while the time of enactment of relational reasoning. Here, human models are attains the 93.1% accuracy ,Q-type base line is 42.5%, LSTM is 48.5%, CNN with LSTM is 53.7%, CNN with LSTM and SA is 69.6%, CNN with LSTM and improved SA is 77.1% and our proposed technique is 98.5% respectively, thus illustration is described in figure 5.

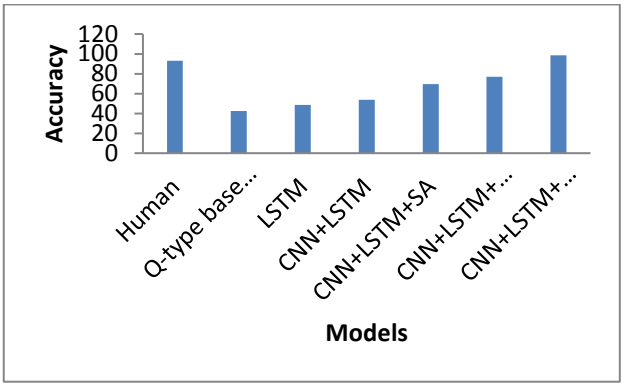

Figure 5: Comparison graph of accuracy with existing techniques

Table 6: Comparison performance of metrics with existing techniques

Models	Data count	Exist data's	Compare Numbers	Query Attribute	Compare Attribute
LSTM	42.5	61.9	70.4	37.4	51.6
CNN+LSTM	44.3	65.8	67.7	49.8	53.6
CNN+LSTM+SA	52.8	71.7	74.1	85.9	52.9
CNN+ LSTM+ improved SA	65.1	83.3	78.1	83.2	76.1
Proposed Technique	92.1	98.3	94.2	98.7	97.8

In table 6 comparison performance metrics for our proposed technique with existing methods. Existing models LSTM, CNN+LSTM, CNN+LSTM+SA, CNN+LSTM+ improved SA and proposed technique data count is 42.5, 44.3, 52.8, 65.1 and 92.1, exist data's are 61.9, 65.8, 71.7, 83.3, and 98.3, compare numbers are 70.4, 67.7, 74.1, 78.1, and 94.2, query attribute is 37.4, 49.8, 85.9, 83.2, and 98.7, compare attribute is 51.6, 53.6, 52.9, 76.1 and 97.8 respectively. Thus it is illustrated in figure 6.

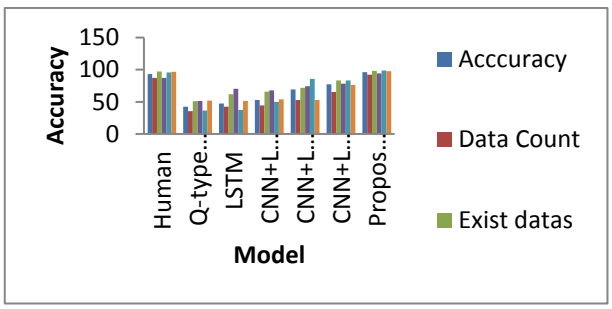

Figure 6: comparison parameter graph for existing and proposed technique

V. Conclusion:

Relevance Networks are efficient, flexible and simple, relationally reasoned neural network modules. Especially notable is the success of RN-augmented networks is state-of - the-art models suggesting that previous architectures lacked a basic, general ability to reason about relationships. In addition, these findings show a significant distinction between the often misunderstood conceptions of perception and reasoning. Powerful visual QA architectures include modules such as ResNets, which are highly capable visual processors that can detect complicated textures and shapes. RNs will easily take advantage of foreknowledge of the relationships to be measured for a specific task. In addition, bounding the otherwise quadratic complexity of the number of relationships could be beneficial, particularly in circumstances with strong computational constraints. Attentive mechanisms may reduce the number of objects fed to the RN as input, thereby reducing the number of relationships to be considered. Our results show that there is no need to cleverly pre-facture the collection of objects, strikingly. RNs learn to deal with "object" representations generated by CNNs and LSTM's. Thus it by probably through the gradients they transmit affecting the content and shape of the representations of objects.

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] Marta Garnelo, Kai Arulkumaran, and Murray Shanahan, "Towards deep symbolic reinforcement learning", *arXiv:1609.05518,* 2016.

[4] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwi´nska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al., "Hybrid computing using a neural network with dynamic external memory", *Nature,* 2016.

[5] Stevan Harnad, "The symbol grounding problem", *Physica D: Nonlinear Phenomena,* vol.42, no.13, pp.335–346, 1990.

[6] Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun, "Tracking the world state with recurrent entity networks", *In ICLR,* 2017.

[7] 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.

[8] 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.

[9] Kushal Kafle and Christopher Kanan, "An analysis of visual question answering algorithms", *arXiv:1703.09684,* 2017.

[10] Charles Kemp and B. Joshua Tenenbaum, "The discovery of structural form", *Proceedings of the National Academy of Sciences*, Vol.105, No.31, pp. 0687–10692, 2008.

[11] M. Brenden Lake, D. Tomer Ullman, B. Joshua Tenenbaum, and J. Samuel Gershman, "Building machines that learn and think like people", *arXiv:1604.00289,* 2016.

[12] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning", *Nature,* vol.521, no.7553, pp.436–444.

[13] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel, "Gated graph sequence neural networks", *ICLR,* 2016.

[14] Mateusz Malinowski and Mario Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input", *In NIPS,* 2014.

[15] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz, "Ask your neurons: A deep learning approach to visual question answering", *arXiv:1605.02697,* 2016.

[16] Allen Newell, "Physical symbol systems", *Cognitive science,* vol.4, no.2, pp.35–183, 1980.

[17] 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.

[18] M. Henaff, J. Weston, A. Szlam, A. Bordes, and Y. LeCun, "Tracking the world state with recurrent entity networks", *arXiv preprint arXiv:1612.03969,* 2016.

[19] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning", *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2017, pp. 2901-2910.

[20] B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Inferring and executing programs for

visual reasoning", *In Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp.2989-2998.

[21] K. Kafle, and C. Kanan, "An analysis of visual question answering algorithms", *In Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp.1965-1973.

[22] B.M. Lake, T.D. Ullman, J.B. Tenenbaum, and S.J. Gershman, "Building machines that learn and think like people", *Behavioral and brain sciences*, vol.40, 2017.

[23] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A deep learning approach to visual question answering", *International Journal of Computer Vision*, vol.125, no.1-3, 2017, pp.110-135.