International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 03 | Mar 2022 www.irjet.net p-ISSN: 2395-0072

A Literature Survey on Image Linguistic Visual Question Answering

[1]Charu Sheela, [2]Hiya Namdeo, [3]Siddhartha Jain, [4]Yash Soni, [5]Prof. Akshatha G

[1]*Dept. of Computer Science and Engineering, Dayananda Sagar College of Engineering, Bengaluru, Karnataka, India*
[2]*Dept. of Computer Science and Engineering, Dayananda Sagar College of Engineering, Bengaluru, Karnataka, India*
[3]*Dept. of Computer Science and Engineering, Dayananda Sagar College of Engineering, Bengaluru, Karnataka, India*
[4]*Dept. of Computer Science and Engineering, Dayananda Sagar College of Engineering, Bengaluru, Karnataka, India*
[5]*Dept. of Computer Science and Engineering, Dayananda Sagar College of Engineering, Bengaluru, Karnataka, India*

Abstract - *VQA is a task that for given text-based questions about an image the system needs to infer the answer for each question, by picking an answer from multiple choices. Many of the VQA systems that have lately been developed contain attention or memory components that facilitate reasoning. This paper aims to develop a model that achieves higher performance than the current state-of-the-art solutions. Also, this paper questions the value of these common practices and aims to develop a simple alternative. We will be exploring the different existing models and developing a custom model to overcome the shortcomings of the existing solutions. We will be benchmarking different models on the Visual7W dataset.*

Key Words: **Computer Vision, Convolutional Neural Network, Natural Language Processing, Visual Question Answering, Artificial Intelligence**

1.INTRODUCTION

COMPUTER VISION

It's a sort of AI in which computers can look at the world, analyze visual data, and then make choices or understand about the environment and scenario.

Computer vision techniques can be employed to train a computer to perform equivalent tasks in much less time by using cameras, data, and algorithms. It can quickly surpass people in finding flaws or issues that cannot be easily identified by humans.

Computer vision is useful for a wide range of sectors, including power and utilities, production, and automobiles, and the field is continually expanding.

COMPUTER VISION: HOW DOES IT WORK?

Computer vision essentially needs a considerable quantity of data to process it a number of times, and thus recognise pictures. In order to train a computer, for documents in order for it to comprehend the distinctions and recognise a tyre, especially one that is free of flaws. A CNN and deep learning, a kind of machine learning, are two fundamental technologies used to do this. A computer can use algorithmic models to teach itself in the context of visual input with the help of Machine learning technology. If the model receives enough data, the computer will "see" it and learn to discriminate between pictures.

A CNN assists a machine learning or deep learning model in "seeing" by fragmenting the pictures into pixels that are assigned labels. Labels are used to execute convolutions to make estimations regarding what is seen. The neural net. perform convolutions and then the precision of the predictions are evaluated till the time these predictions prove to be correct. Then it identifies or perceives pictures in a similar way to a human.

A CNN recognises hard edges and simple shapes first, then fills in the details as it performs iterations of its predictions, much like a person recognising a picture from a distance. A CNN is used to comprehend single pictures.

VISUAL QUESTION ANSWERING (VQA)

It's an intriguing training environment for assessing the capabilities and flaws of present picture comprehension algorithms. Many of the VQA systems that have lately

been developed contain attention or memory components that facilitate "reasoning." Nearly all of these systems use picture and question data to train a multi-class classifier to determine a response in multiple-choice VQA.

1.1 Background and Motivation

VQA appears to be an ideal environment for developing systems that can conduct rudimentary "reasoning" about images. A number of researchers have lately looked into adding basic memory or components based on attention to VQA systems.

While these systems have the ability to do simple reasoning in theory, it is unclear whether they actually do so or in a fashion that is understandable to humans. This research aims to answer a basic question: are these systems superior to baselines that are only meant for example, to recognise car tyres, it needs to be fed a vast number of tyre photographs and tyre-related capture the dataset bias of normal VQA datasets? We narrowed the application of our research to multiple-choice tasks in order to conduct a much more experimental investigation free of the complexities of assessing produced material.

2. LITERATURE SURVEY

[1] Jia, P., et al.,: Going deeper with convolutions.
A deep convolutional neural network architecture by the name of Inception is proposed. It establishes a new standard for detection and classification (ILSVRC14). The most basic characteristic of this particular architecture is the increased usage of resources of computing inside the network. Even inside the convolutions, the essential approach to do this is to move from fully connected to sparsely connected structures. State-of-the-art sparse arithmetic operations result from grouping sparse matrices into comparatively dense submatrices. Increasing the scale of deep neural networks is the most simple method for improving their performance. This includes increasing the depth and the breadth of the network. Another advantage of this method is that it is based on the idea of processing the visual input to numerous scales before being aggregated, allowing the following step to abstract data from multiple scales at once.

[2] Parikh, Lu, et al.,: VQA: Visual question answering.

A VQA system is provided with an image as input and answers in natural language about the query. This goal oriented exercise can be used in situations where visually challenged users or intelligence analysts are actively eliciting visual data. The AI capabilities required to answer open-ended questions include object detection, fine-grained recognition, reasoning based on knowledge, recognising tasks, and commonsense understanding. This publication includes an open-ended as well as a multiple-choice problem. It includes a big dataset with 204,721 pictures from the MS COCO dataset as well as a freshly constructed abstract scene dataset with 50,000 scenes. The MS COCO dataset contains photos displaying a wide range of scenarios that are effective in evoking a wide range of questions.

[3] Zemel, et al.,: Exploring models and data for image question-answering.

This article explains its contributions to the problem, which includes a generic complete QA model that connects a Convolutional Neural Network and a recurrent neural net (RNN) using visual semantic embeddings, as well as analogies to a number of other models; an automated question generation algorithm which transforms descriptive sentences into questions; and a fresh QA dataset (COCO-QA) that was created using the algorithm, as well as a number of baseline results on the newly created dataset. The replies are considered to be single words in this work, allowing the topic to be treated as a classification problem. This also simplifies and strengthens model evaluation, eliminating the vexing evaluation concerns that afflict multi-word generation problems. This model outperforms the only reported findings on an existing picture QA dataset by 1.8 times. Because the current dataset is inadequate, an algorithm was developed to aid in the collecting of a large-scale picture quality assurance dataset using image descriptions. The question-generating method may be used on a variety of picture-description datasets and can be automated with minimal human intervention. Image question answering is a very young topic of research, and the approach utilized here has a number of flaws. These models are just answer classifiers, to begin with.

[4] Berg, T., et al.,: Visual madlibs: Fill in the blank image generation and question answering.

The Visual Madlibs dataset was built by collecting specific descriptions of objects and people, their relationships, activities and appearance, including assumptions about the complete context or scene, using automatically generated fill-in-the-blank templates. Many experiments have been conducted on the Visual Madlibs dataset, and its use in two unique description generating tasks has been demonstrated: focused description development and multiple-choice question answering for images. Using this new dataset, this

research develops methods for producing more focused descriptions. These kinds of descriptions aim for high level aims like creating human-like visual interpretations of images.

[5] Groth, et al.,: Visual7w: Grounded question answering in images.

The 7W questions, which are visually based, are proposed in this study as a way to help individuals grasp pictures more profoundly than merely recognising things. Previous attempts to build a strong semantic link between written descriptions and picture areas failed miserably. The bounding boxes in the photos are linked to the item mentioned. Using object level grounding, a semantic relationship is created between regions of the image and descriptive text. This paper resolves coreference ambiguity, comprehends object distributions, and evaluates on a new kind of visually grounded QA using object grounding. An attention-based LSTM model is recommended for achieving advanced performance on QA tasks. Visual QA tasks are explored in a grounded context using a big collection of 7W multiple-choice QA pairings. On the QA tasks, human performance as well as many baseline models are also examined. Finally, in order to solve the 7W QA problems, a unique LSTM model incorporating spatial attention is offered. This study provides a dataset that expands on previous research and a model for doing this job based on attention.

[6] J. Yang, et al., "Hierarchical question-image co-attention for visual question answering,"

For visual question answering, the model that has been proposed is hierarchical in nature. This model can attend to diverse areas of image, including various components of the query. It's equally important to demonstrate which words to pay attention to as it is to model where to look for visual attention. To capture information from multiple granularities, the question is modeled hierarchically at three levels. A novel VQA co-attention model is introduced that combines image and question attention reasoning. Furthermore, this model uses a unique 1-dimensional convolution neural network to reason about the query in a hierarchical approach (CNN). On the VQA dataset, the approach improves from 60.3 percent to 60.5 percent, and on the COCO-QA dataset, it improves from 61.6 percent to 63.3 percent. The performance of VQA and COCO-QA is enhanced to 62.1 percent and 65.4 percent, respectively, when ResNet is used. T This strategy was only tested on visual question answering but it'll suffice to say that it can be used to address other language and vision difficulties.

[7] Neil Hallonquist, et al., "Visual Turing test for computer vision systems"

This work attempts to create a novel query-based computer vision test, which is one of the most active areas of modern AI research. Throughout this work, the phrases "computer vision" and "semantic picture interpretation" are being used consistently. The goal of this study is to develop a quantitative measure of how well a computer vision system can understand common photographs of natural scenes. Although it concentrates on urban street scenes, the underlying logic and motives can simply be applied to other picture populations. The vocabulary V in this study is made up of three parts: kinds of items (T), type-dependent properties of objects (A), and type dependent interactions between two objects (V). Each inquiry $q \in Q$ falls into one of four categories: existence, Qexist; uniqueness, Quniq; attribute, Qatt; and relationship, Qrel.

[8] Mengye Ren., et al., "Image Question Answering: A Visual Semantic Embedding Model and a New Dataset,"

Instead of object recognition and picture segmentation as intermediary steps, this study recommends employing RNN and visual semantic embeddings. The problem of interest is learning image and text simultaneously through a question-and-answer activity. Recurrent neural networks and visual semantic embeddings reused in the model, and a newly proposed method of synthesizing QA pairings from the currently available picture description dataset is used in the dataset. For DAQUAR, the best model employs a 300-dimensional embedding that is randomly initialized, but for COCO-QA, the best model uses a problem-specific skip-gram embedding. Fine-tuning word embedding and normalizing CNN hidden image characteristics were revealed to boost performance. Despite having a good comprehension of the subject and some coarse picture awareness, this model remains naïve in many cases. The question-generating approach may be used for a variety of picture description datasets and can be automated without a lot of human interaction.

[9] Wei Xu Baidu, et al., ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering,

For VQA challenges, this study presents a new attention-based deep learning architecture. VQA delivers a natural language answer when given an image and a query about it. VQA helps blind people with image representation, primary learning, and navigation.An

RNN model has been used to anticipate the future attention area based on the present attention zone's geographical & visual attributes. The ABC-CNN framework for the VQA problem is proposed in this paper. Through the use of a question-guided attention map, this paradigm integrates visual feature extraction and semantic question comprehension. A programmable convolution network generates the attention map, which is adaptively defined by the meaning of questions. Both visual question responding capability & awareness of the combination of question semantics & picture contents are considerably improved by ABC-CNN.

[10] Lin Ma, et al., "Learning to Answer Questions From Image Using Convolutional Neural Network,"

This study suggests the use of CNN for visual question answering (VQA). The paper provides an end-to-end framework by making use of the CNN for learning both the inter-modal interactions and the picture and question representation for generating the response. In order to generate the answer, it is necessary to learn not just the picture and question representations, but also their intermodal interactions. The model consists of three CNNs: an image CNN for encoding picture data, a sentence CNN for composing question words, and a multimodal convolution layer for learning their combined depiction for classification in the space of candidate answer words. This CNN model uses the convolutional architectures for creating picture representations of picture, composing words that are consecutive according to the representation of question, and learning the relationships between question and image predicting the response.

[11] Saurabh Gupta, et al.,, "Exploring Nearest Neighbor Approaches for Image Captioning,"

For picture captioning, the paper investigates a range of closest neighbor baseline techniques. These methods use the training set to find a set of nearest neighbor photos from which a caption for the query image can be borrowed. It selects a caption for the query picture by assessing whichever caption best reflects the "consensus" of candidate captions obtained from the query image's closest neighbors. When evaluated using automatic assessment metrics on the MS COCO caption evaluation server, these techniques surpass several current algorithms that create novel captions. Human studies, on the other hand, suggest that a method that creates original captions is still preferable to the nearest neighbor method. The nearest neighbor method returns a set of k photos. To identify which photographs are "nearest," GIST, pre-trained deep characteristics, and deep characteristics fine-tuned for the purpose of caption development are all employed. Once a collection of k NN photographs has been located, the captions explaining these images are gathered into a set of selected captions from which the final caption is picked. It chooses the best candidate caption by looking for the one with the highest score when compared to the others. The "consensus" caption is what it's called. The CIDEr or BLEU metrics are used to calculate the scores between caption pairings.

[12] D. Chen and C. D. Manning, "A fast and accurate dependency parser using neural networks.,"

By combining current breakthroughs in natural language processing and computer vision, we offer a method for automatically answering queries regarding images. A multi world method, which reflects uncertainty about the seen world in a Bayesian framework, combines discrete reasoning with uncertain predictions. The method can answer sophisticated human questions regarding realistic scenarios with a variety of responses like counts, object classes, instances, and lists of them. Question-and-answer pairs are used to directly train the system. We create a first benchmark for this activity, which may be thought of as a modern take on a visual turing test.

[13] Mario Fritz, et al., "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input"

By combining current breakthroughs in natural language processing and computer vision, we offer a method for automatically answering queries regarding images. In a Bayesian framework, we integrate discontinuous reasoning with indeterminate predictions using a multiworld approach to explain uncertainty about the seen world. The method can answer sophisticated human questions regarding realistic scenarios with a variety of responses like counts, object classes, etc. Question-and-answer pairs are used to directly train the system. We create a first benchmark for this activity, which may be thought of as a modern take on a visual turing test.

[14] Mateusz Malinowski, et al., Ask Your Neurons: A Neural-based Approach to Answering Questions about Images

We tackle a Visual Turing Test-style question answering assignment using real-world graphics. Through combining previous accomplishments in NLP with input images, we introduce Neural-Image-QA, an end-to-end solution to this question in which all aspects are simultaneously taught. With the exception of previous

attempts, we are faced with a multi-modal problem in which natural and visual language input impact linguistic outputs (reaction) (question and image). The Neural-Image-QA technique outperforms the prior best response towards this topic by a factor of two. By assessing how much data is just there in the language part, we contribute to our knowledge of the situation and establish a new human baseline. We propose two unique measures and gather extra answers to explore human consensus, which is connected to the uncertainties inherent in this difficult activity, extending the original DAQUAR dataset to DAQUAR Consensus.

[15] Min-Oh Heo, et al., "Multimodal Residual Learning for Visual QA"

With diverse methodologies, deep neural networks continue to progress in image identification problems. However, there are few applications of these strategies to multimodality. This study offers Multimodal Residual Networks (MRN), which expand the concept of deep residual learning to multimodal residual learning of visual question-answering. The basic idea is to employ element-wise multiplication for joint residual mappings, which takes advantage of current research on attentional model residual learning. Based on our research, various alternative models presented by multimodality are investigated. Furthermore, we provide a novel back-propagation approach for visualizing the attention effect of joint representations for each learning block, despite the visual features being collapsed without spatial information

[16] Haoyuan Gao, et al., "Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering"

The paper presents the mQA system that could solve questions about the content about an image. A sentence, a phrase, or a single word might be used as an answer. A CNN extracts the visual representation, a Long Short Term Memory extracts the question representations, and LSTM saves the linguistic context in a solution. The fusing element fuses the data from the former three elements to produce the solution. For training and evaluating the mQA model, a Freestyle Multilingual Image Question Answering dataset was created. There are around 150 thousand images and 310 thousand freestyle Chinese question-answer combinations with English language transcriptions in the collection. Towards this dataset, human judges apply a Turing Test to evaluate the accuracy of the mQA model's produced replies. In particular, we combine human responses with our model. The human reviewers should be able to distinguish amongst the model and the real thing. They'll also give the answer a score (i.e.To express how good it is, use the numbers (0, 1, 2, the greater the number, the better). We recommend methods for monitoring the quality of the assessment process. Human judges cannot identify our model from humans in 64.7 percent of cases, according to the experiments. The overall average is 1.454. (1.918 for humans).

[17] Kaii Chenn, et al., "Efficient Estimation of Word Representations in Vector Space"

In this paper unique model architectures are presented for developing non-discrete word vectors from massive datasets. In a word similarity task, the accuracy of such depictions is tested, and the results have been compared to past top-performing systems depending on multiple types of neural networks. Significant gains have been taken under consideration, in accuracy at a fraction of the cost of computing; for eg. - producing high-quality word vectors from a 1,600,000 word given dataset takes approximately 24 hours. State-of-the-art results were found to be produced by these vectors, while analyzing semantic and syntactic word similarity on the set of tests.

[18] Marshall R. Mayberry, et al., "SARDSRN: A Neural Network Shift-Reduce Parser"

Chronic pain patients frequently feel that their discomfort is caused by the weather. Scientific evidence to back up these claims is unclear, in part because it's difficult to obtain a large dataset of patients who report their pain symptoms frequently and in a range of weather situations. Smartphones make it possible to collect data in order to overcome these obstacles. Our research, Cloudy with a Chance of Pain, looked at daily data from 2658 patients over the course of 15 months. The researchers discovered significant but small links amongst pain and wind speed, relative humidity, and pressure with associations remaining even after mood and physical exercise were accounted for. This study demonstrates how citizen-science studies can be used to collect massive datasets from real-world populations in order to answer long-standing health problems. These findings will serve as a foundation for a future system that would help patients better manage their health by providing pain forecasts.

[19] D. McClosky, et al., "The stanford corenlp natural language processing toolkit.,"

This article describes the Stanford CoreNLP toolbox, an expandable pipeline that provides core natural language analysis. This toolbox is widely used in the NLP academic community, as well as business and

government NLP users. This, we feel, is attributable to a simple, approachable design, simple interfaces, the inclusion of reliable and high-quality analytical components, and the lack of a lot of baggage. This article describes the design and development of Stanford CoreNLP, a Java annotation pipeline system that incorporates the majority of the typical core natural language processing (NLP) operations, from tokenization through coreference resolution.

[20] Y. Yu,et al., "Tensorflow: A system for large scale machine learning,"

TensorFlow is a ML framework. In a dataflow graph, its nodes are distributed among several systems in the cluster. It also provides diverse processing units including CPUs, GPUs, and Tensor Processing Units, which are custom-designed ASICs (TPUs). TensorFlow's design allows application developers to use new optimizations and training strategies, whereas previous "parameter server" systems integrated shared state management into the system.

[21] Razavian, A.S., et al., CNN features off-the shelf: an astounding baseline for recognition.

The OverFeat neural network was trained using the ILSVRC13 dataset to perform object classification, scene recognition, attribute detection, and picture retrieval. Except for the Sculpture dataset, it consistently beats minimal memory footprint approaches in retrieval. Deep learning using CNN appears to be the clear choice in many of the visual recognition activities, according to the statistics. Object detection does not necessitate object localization. For the ILSVRC object image classification job,the CNN representation has been optimized.

[22] Kim, J., et al., Multimodal residual learning for visual QA.

Multimodal Residual Networks (MRN) are being developed for visual question-answering residual learning that goes beyond deep residual learning. MRN uses element-wise multiplication to learn joint representation from visual and verbal data. The research produces excellent results on the VQA dataset using the back-propagation strategy for both Open-Ended and Multiple Choice tasks. Deep residual learning is a framework for deep neural networks that improves object recognition studies while also giving a broad paradigm. Another projection of function F(x), a remainder of the identity mapping of x, is fitted using the nonlinear layers of neural networks already in place. In this technique, very deep neural networks are able to efficiently learn representations. Furthermore, it may aid in the resolution of real-world problems that necessitate the use of an integrated approach.

[23] Mikolov, T., et al., Efficient estimation of word representations in vector space.

For creating continuous vector representations of words from very large data sets, 2 novel model architectures are given. The quality of all these depictions is examined in a word similarity challenge, and the comparison is made to the best and most consistent algorithms depending on multiple neural network models. It demonstrates considerable gains in accuracy at a minimal computational cost. Many modern NLP systems and methodologies regard words as atomic units, with no concept of word similarity because they are represented as indices in a dictionary. Simple models that have been trained on a large amount of data often outperform complicated systems that have been trained on a smaller amount of data.Within several cases, just upgrading fundamental procedures won't yield substantial results, and we must instead emphasize on the more complex strategies.

[24] Parikh, J., et al., Hierarchical question-image co-attention for visual question answering.

A large number of studies have presented attention models for Visual Inquiry Answering (VQA) that produce spatial maps identifying picture areas connected to the query. In this research, we propose that modeling "what words to listen to" or questioning attention is equally significant. We propose a novel VQA co-attention model that reasons over images and questions attention using a novel 1-d convolutional neural network (cnn). In both academia and industry, visual question answering (VQA) has emerged as an important multi-discipline challenge. The system must comprehend both the image and the inquiry in order to appropriately answer visual inquiries about it. We suggest in this research that determining "which words to listen to" or questioning attention is just as crucial. We're going to show you something new.

[25] Fukui, A., et al., Multimodal compact bilinear pooling for visual question answering and visual grounding.

To Combine Multi-modal features economically and expressively, bilinear pooling (MCB) is used. On the visual question responding and grounding tasks, MCBoutperforms ablations without MCB. This paper

recommends employing Multimodal Compact Bilinear pooling (MCB) to produce a shared representation. MCB is used to compute the outer product of two vectors, allowing for a multiplicative interaction among all elements of both vectors. In vision-only testing, it has been shown to aid fine-grained categorization. Image and text are projected randomly to estimate Multimodal Compact Bilinear pooling (MCB) to a higher dimensional space.

TABLE 1: ANALYSIS AND REMARKS ON THE CURRENT STUDY

Author	Year	About the paper	Advantages	Disadvantages
Zemel, et al.	2015	This article explains its contributions to the problem, which includes a generic complete QA model, as well as analogies to a number of other models; an automated question generation algorithm; and a fresh QA dataset (COCO-QA) that was created using the algorithm, as well as a number of baseline results on the newly created dataset.	• Elimination of intermediate stages like image segmentation and object detection. • 1.8 times better performance compared to the only available dataset. • Robust and easier evaluation.	• Cannot allow longer answers. • limited set of questions. • tough to interpret the reason behind a certain output. • Lacks in visual attention.
Berg, T., et al.	2015	This research develops methods for producing more focused descriptions which have high-level aims like creating human-like visual interpretations of images.	• Focused natural language description generation task. • Multiple-choice question answering task for images. • fill-in-the blank strategy.	• Not very useful for fine-grained questions. • Does not give better accuracy when collecting visual characteristics for other questions.
J. Yang, et al.	2017	A novel VQA co-attention model is introduced that combines image and question attention reasoning.	• Performance improvement from 60.3% to 60.5%, on the VQA dataset and from 61.6% to 63.3% on the COCO-QA dataset. • Good performance for questions having a limited set of answers. • Defines 3 levels in architecture: word, phrase and question level.	• Does not perform very well for questions that have many different answers. • Cannot identify many complex patterns.

Mengye Ren. et al.	2015	Instead of object recognition and picture segmentation as intermediary steps, this study recommends employing RNN and visual semantic embeddings.	• Demonstrates decent understanding of the question as well as some basic image interpretation • The algorithm helps in collection of large-scale QA dataset. • Extensible question generation algorithm.	• Model is merely an answer classifier. • Proposed algorithm assumes all responses are single word answers. • Algorithm implementation is question type dependent.
Kan Chen. et al.	2016	This study presents a new attention-based deep learning architecture. VQA delivers a natural language answer when given an image and a query about it.	• Improved VQA ability. • Provides better understanding of how question semantics and image contents are linked. • Capable of providing more appropriate results.	• New research shows simpler models like binary classifiers perform equivalent to proposed solutions. • Does not mention any scope for future enhancements.
Jacob Devlin, et al.	2015	For picture captioning, the paper investigates a range of closest neighbor baseline techniques.	• Demonstrates a simple NN approach providing a baseline for image captioning. • Offer outcomes for a number of different NN approaches.	• Approaches based on NN can result in fairly generic captions. • When undertaking human evaluation, it misses nuances in the likeness and contrasts among the systems.
Danqi Chen, et al.,	2015	This study used neural networks to create a unique dependency parser.	• In terms of accuracy and performance, the parser surpasses existing greedy parsers that use sparse indication characteristic • Cost effective feature computation. • Able to automatically learn feature conjunction for predictions.	• Lacks in the accuracy of the neural net. • Room for improvement in the architecture.

Richard Socher, et al.,	2014	It is a novel universal log bilinear regression model for unsupervised word representation learning that outperforms previous models on tasks including word similarity, word analogies, and named entity identification.	• Count-based approaches' accuracy in capturing statistical information might be useful. • The model efficiently uses statistical data.	• When the number of negative samples exceeds roughly 10, the performance of word2vec really degrades.
Andrej Karpathy, et al.,	2014	This research offers a novel dependency tree-based recurrent neural network.	• It provides more comparable representations to phrases describing the same visual. • Many recurrent and recursive neural networks are outperformed by DT-RNNs. • It is sturdier in comparison to other neural-net models.	• The SDT RNN's fails when a phrase that describes a picture does not have a verb, but the other sentences in that image do.

3. CONCLUSION

As a result of the above survey on the Visual Question Answering (VQA), it is evident that there are still some issues that need to be addressed. Until recently, the research has suggested featurizing the responses and training a binary classifier to predict the validity of a triplet of picture questions and answers. Though the models perform well there is still the scope of improvement and the ideas lack clear implementation. This provided the impetus for our research and development of a better alternative. The proposed point and then look at tweaking the model to improve its performance. Experiments on the real-world Visual7W dataset have demonstrated the efficacy of co-attention models, particularly when image question-linguistic co-attention is considered.

REFERENCES

[1] Liu, W., Jia, Y., Erhan, Sermanet, P., Reed, Szegedy, C., S., Anguelov, D., Rabinovich, D., Vanhoucke, V., A.: Going deeper with convolutions.

[2] Antol, S., Lu, J., Agrawal, Parikh, D, A., Mitchell, M., Batra, D., Zitnick, C..: VQA: Visual question answering. In: Proceedings of the International Conference on Computer Vision.

[3] Ren, M., Zemel, R., Kiros: Exploring models and data for image question answering. In: Advances in Neural Information Processing Systems.

[4] L., Yu, E., Berg, A., Berg, Park, T.: Visual madlibs: Fill in the blank image generation and question answering.

[5] Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images.

[6] Lu, D. Batra, D. Parikh and J. Yang,, Hierarchical question-image co-attention for visual question answering, in Advances In Neural Information Processing Systems

[7] Donald Gemana, Neil Hallonquist, Stuart Gemanb "Visual Turing test for computer vision systems"

[8] Ryan Kiros, Richard Zemel., Mengye Ren, "Image Question Answering: A Visual Semantic Embedding Model and a New Dataset,"

[9] Haoyuan Gao, Jiang Wang, Kan Chen, Wei Xu Baidu, Ram Nevatia, Liang-Chieh Chen, "ABC CNN: An Attention Based Convolutional Neural Network for Visual Question Answering,"

[10] Qi Wu, Peng Wang, Chunhua Shen, "Image Captioning and Visual Question Answering Based on Attributes and External Knowledge,"

[11] Ross Girshick, Saurabh Gupta, Jacob Devlin, "Exploring Nearest Neighbor Approaches for Image Captioning,"

[12] C. D. Manning and D. Chen, "A fast and accurate dependency parser using neural networks.,"

[13] Mario Fritz and Mateusz Malinowski, "A Multi World Approach to Question Answering about Real-World Scenes based on Uncertain Input"

[14] Marcus Rohrbach, Mateusz Malinowski and Mario Fritz, "Ask Your Neurons: A Neural-based Approach to Answering Questions about Images"

[15] Jin-Hwa Kim, Dong-Hyun Kwak, Sang-Woo Lee and Min-Oh Heo, "Multimodal Residual Learning for Visual QA"

[16] Jie Zhou, Zhiheng Huang, Haoyuan Gao, Junhua Mao, Lei Wang and Wei Xu1, "Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering"

[17] Kai Chen, Tomas Mikolov, Jeffrey Dean and Greg Corrado, "Efficient Estimation of Word Representations in Vector Space"

[18] Marshall R. Mayberry, III and Risto Miikkulainen, "SARDSRN: A Neural Network Shift-Reduce Parser"

[19] J. R. Finkel, C. D. Manning, J. Bauer, M. Surdeanu, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit.," in ACL (System Demonstrations)

[20] M. Abadi, D. G. Murray, P. Barham, J. Chen, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, Z. Chen, A. Davis, J. Dean, M. Devin, M. Wicke, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, Y. Yu, and X. Zheng, "Tensorflow: A system for large scale machine learning,"

[21] Sullivan, Azizpour, Carlsson, Razavian, J.S., A.S: CNN features off-the-shelf: an astounding baseline for recognition. In: arXiv 1403.6382.

[22] Kim, J., Kwak, Zhang, Lee, S. D., J., J., Heo, M., Kim, Ha, B.: Multimodal residual learning for visual QA.

[23] Mikolov, Corrado, G., Dean, J.Chen, T., K.,: Efficient estimation of word representations in vector space.

[24] Lu, Yang, Parikh, J., Batra, D., J., D.: Hierarchical question-image co-attention for visual question answering.

[25] D., Rohrbach, A., Yang, D., Darrell, A. Fukui, T., , Huk Park, Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding.