Visual Question Answering using Deep Learning

Pallavi K S[1], Sonali M A[2], Tanuritha P[3], Vidhya L[4], Prof. Anjini L[5]
[1,2,3,4] *Student, Department of Computer Science and Engineering, Bangalore Institute of Technology, Bangalore*
[5] *Assistant Professor, Department of Computer Sciece and Engineering, Bangalore Institute ofTechnology, Bangalore*

Abstract*: This paper introduces the concept of VQA (VisualQuestion Answering), which uses CNN (Convolutional Neural Network) attention model and innovative LSTM (long short-term memory) and CNN (Convolutional neural network) attention models to combine local image features and questions from corresponding specific parts or regions of an image to provide answers to questions posed using a pre-processed image dataset. Here, the word attention can be explained as using techniques which allow the model to only emphasize those elements of the image that are relevant to both the image and the key phrases within the question. The areas of the image that are irrelevant will not be taken into account, improving classification accuracy by lowering the chances of guessing incorrect answers. Use of the Keras Python package with the backend of Tensor Flow, followed by the NLTK Python libraries, for the purpose of extracting image features with the help of CNN, the language semantics with the help of NLP, and finally use of the multi- layer perceptron for the purpose of combining the outcome or results from the question and the image.*

Index Terms— **VQA, CNN, RNN, AI, LSTM, NeuralNetworks, Image Processing**

INTRODUCTION

The field of computer science known as "Visual Question Answering" is focused on the creation of a system that is able to provide responses to questions based on a picture as well as spoken language. The Visual Question and Answering (VQA) task is one that has been fully automated by artificial intelligence. It combines NLP (Natural Language Processing) with computer vision (CV). If we provide our model with an image and a set of questions, it will generate not only a solution but also a collection of explanations (in the form of text) and visual attention maps. Questions pertaining to images that can be answered in any way are included in the VQA dataset. To answer these questions, you'll require vision, language, and common sense. The VQA models will have the capability to reaso006E with the assistance of unstructured external knowledge sources (text detected in a test image), and they will also be able to manage multiple data streams (predicting a solution from a predetermined vocabulary or providing a solution via copy). To launch study and broaden the limits of both disciplines, visual question answering was given as a means to merge computer vision with NLP (Natural Language Processing. Computer vision is the study of techniques for capturing, analysing, and interpreting pictures. In summary, the idea is to teach robots how to sight. Furthermore, NLP is the branch of computer science that studies how computers and humans interact with natural language, such as teaching machines to read. Both computer vision and natural language processing fall under the umbrella of artificial intelligence and use the same machine learning techniques. However, they have had individual disparities. Both professions have witnessed substantial expansion in recent decades, and the combined huge increase of visual and textual data is forcing a convergence of efforts from both domains. A lucrative strategy is to integrate CNNs (Convolutional Neural Networks) trained for object identification with word embeddings developed for big text libraries. The most common kind of Visual Query Answering (VQA) involves providing the computer both a picture as well as a written question pertaining to the image. It is anticipated that the computer would provide the appropriate response, which may come as a few words or a short sentence. The options may include binary settings (yes/no) and multiple-choice settings, in which candidate responses are provided to the user. "Fill in the blank" is a similar game in which you have to finish a sentence about a picture by filling in gaps with the right words.. These phrases are simply inquiries that have been rephrased in the form of declarative statements. One of the most significant distinctions that can be made between VQA and other kinds of work is that the question that has to be answered is not processed until run time. As a result of the fact that VQA often needs information in addition to that which is already there in the picture, it is a challenge that is substantially more difficult to solve than image captioning. The kind of additional information required can vary from common sense to detailed knowledge of specific image elements. In this regard, VQA is a proper AI-complete task, since it requires understanding from multiple subdomains. This encourages the interest in VQA, by demonstrating the development of AI systems having the capability of advanced explanation coupled with deep language and image understanding. Image understanding through image captioning is evaluated equally well. Practically however, the evaluation through VQA is a lot easier giving VQA the advantage. Typically, the answers contain only a few words. It is more hard to differentiate anticipated captions to long ground truth captions. Despite a thorough examination of advanced evaluation metrics, this remains an open research problem. This questionnaire begins with a thorough examination of several VQA methodologies in its very first part. This thorough examination is divided into four categories depending on the nature of their major contribution. The presence of incremental contributions indicates that majority methods fall into more than one of these categories. The upsurge of neural network in natural languages processing sparked the development of joint embedding techniques at first.. They use convolutional and recurrent neural networks, abbreviated as CNNs and RNNs, respectively, in order to learn the embeddings of images and sentences in a shared feature space. As a result, they can be fed into a classifier to anticipate an answer. The next step, called attention mechanisms, improves the process that came before it by concentrating on certain aspects of the information that may be imagined or questioned. The successful implementation of some procedures

within the scope of picture captioning sparked interest in VQA. The main intension is to replace image-wide features with spatial feature maps and to interact with the question and particular regions of these maps. In addition, compositional models attach the computations performed to each problem instance. The second part of this study takes a look at the datasets that are readily available to be used in the training and assessment of VQA systems.

(i) their size

(ii) the amount of required reasoning

(iii)The amount of information required in addition to what is present in the actual images.

With few exceptions, the review highlights that existing datasets rely on visual-level questions and require only a minor percentage of external knowledge. These properties highlight the conflict with simple visual questions that the current state of the art can handle, but they should not be overlooked when Visual Question and Answering is displayed as an AI-complete evaluation alternative. Thus, it can be deduced that datasets which are more diverse and sophisticated will sooner or later will be required. Another major benefaction of this survey is the deep examination of the question/answer pairs present in the Visual Genome dataset. They are included in the most extensive VQA dataset that is currently accessible, which is rich in structured picture annotations and is formatted in the form of scene graphs. The usefulness of these annotations for VQA is established by comparing the development of principles in the questions, responses, and picture annotations given. It has been revealed that only around forty percent of the responses are comparable to the characteristics shown in the scene graphs. It is further demonstrated that by connecting scene graphs to external knowledge sources, this similarity may be significantly increased. Section 5 concludes this study by examining the possibility of expanded connectivity to such knowledge sources, as well as enhanced utilization of existing work in the field of NLP. This concludes the story.

LITERATURE REVIEW

1. P. Gao [2019]: The work emphasizes few frequently used deep learning architectures along with their applications. The networks that are given in depth are the CNN (Convolutional Neural Network), the auto encoder, the deep belief networks, Boltzmann machine. Deep learning can be used to process data in unsupervised learning algorithms.
2. L.Ma[2016]: The suggested technique is to employ convolutional neural network to learn to create responses queries relevant to a picture. This section provides an overview of various aspects of Visual Question Answering and algorithms.
3. J. Pennington [2014]: Glove: Wordrepresentation using global vectors describes techniques
4. for classification such as Convolutional Neural Networksand how to determine its rule.
5. There have been quite a few approaches available to counter the challenge of VQA, mainly by using Artificial Neural Networks particularly Recurrent Neural Network (Iqbal Chowdhury et.al) and Convolutional Neural Network (Qi Wu 2017).

Comparison of various strategies based on test accuracies of answering the question accurately is carried out. According to the findings, primary methods such as (Yuetan Lin et.al.2016) were based on a image of smallersize and a simpler question dataset. Models that were tested using external input.

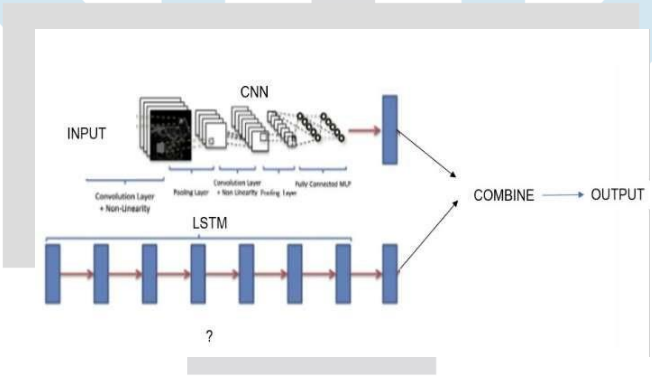


Figure 1: System Architecture of VQA

Information (Qi Wu 2017) produced better results and proved to be a better approach. The approaches were not solely based on the provided image data and a question- answer pair related to the image, but also included extra information about the image that explained in detail about the image's available aspects. Furthermore, attention- based models (Peter Anderson 2018) were developed. The above models emphasize on the image set's feature extraction phase.

METHODOLOGY

Putting forward the concept of a VQA System that offers a comprehensive comprehension of images via the use of fine-grained analysis. When given an image and a question about it, the job is to deliver a response by interpreting the image. Visual questions target different part of the image including the background of the image. As a result, need more thorough understanding of the image. Both a bottom-up and a top-down attention mechanism are being implemented during the course of this Endeavour. The bottom-up method is founded on Faster R-CNN (Regions with Convolution Neural Network), while the top-down mechanism is founded on LSTM, which is a subset of RNN. Both methods are based on RNN. By using this

approach, are able obtain abetter understanding of an image. Able to solve the VQAchallenge and also get better efficiency. The task of VQAinvolves understanding the elements of the images. Oftenprior non-visual information that ranges from "common sense" knowledge which can be encyclopedic or based onthe topic is required. Faster Region-based-CNN (RCNN)is utilised for extracting picture elements with an additional fully connected layer whose weights are dynamically acquired by LSTMs cell according to the inquiry, efficiently dealing with various types of inquiriesthat need variable levels of semantic comprehension. Onthe MSCOCO 2014 (Tsung-Yi Lin et al) training dataset of 82000 pictures, VGG networks and LSTM achieved a training accuracy of 60%. This model achieved a testing accuracy of 57% when compared to the same dataset. With RSVQA, a system can be proposed to get data from remote sensing data which is available to each user: questions are constructed in natural language and are utilized to communicate with the images. A natural language answer is generated in order to extend the approach of using cartoon imagery to real world images.

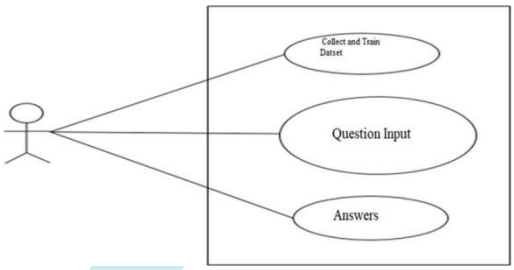


Figure 2: Use case diagram of VQA

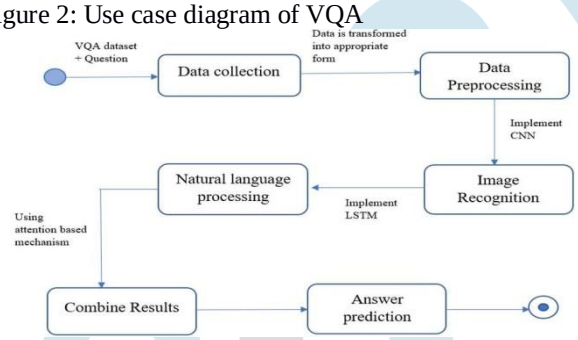


Figure 3: State chart diagram of VQ

RESULTS

Figure 5 shows the VQA page where in we canchoose an image from a set of images and ask a natural language question related to the image and then receive a set of possible natural language answers.

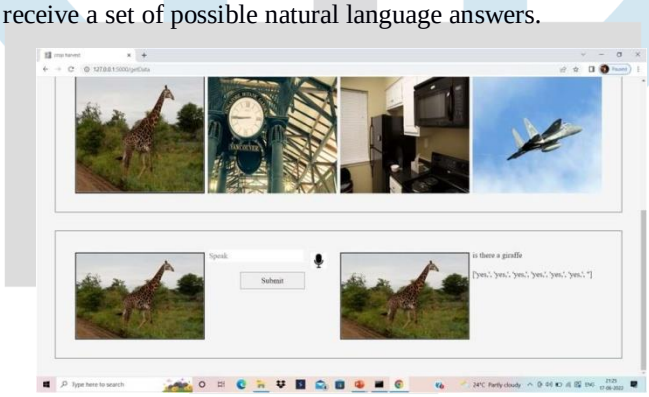


Figure 6: Snapshot of output

The Figure 8.4 shows the snapshot of the output, which is a list of possible natural language answers to the asked question.

APPLICATION

A majority of attention-based deep neural networks are used for VQA. Top-down methodologies are frequently used to classify these models, with context given by a representation of the question in the case of VQA. By anticipating a weighting for each spatial position included within the CNN output, attention may be given to the output of one or more layers of a Convolutional neural network (CNN). In any case, deciding the optimal number of image districts constantly requires an unwinnable exchange off among coarse and fine degrees of detail. Inaddition, the arbitrary positioning of the area relative to the image content may make it more difficult to detect objects that are poorly aligned with the area and to bind visual concepts associated with the same object. Relatively speaking, previous work rarely considers focusing on the salient image area. Coming to the conclusion that there are two papers, Jin et al. employs selective search to locate important picture regions, which are then filtered by a classifier, scaled, and CNN encodedas input to the image caption model. This process is discussed in the previous sentence. In the course of this effort, rather of relying on hand-crafted or differentiable region suggestions, make use of Faster R-CNN. Preparedto do a pretraining session on object identification datasets with this methodology for our area bids.

Conceptually, the benefits should be almost like pre- training visual representations on ImageNet and leveraging significantly larger cross-domain knowledge. Applying the method to VQA, establishing the wideapplicability of the approach.

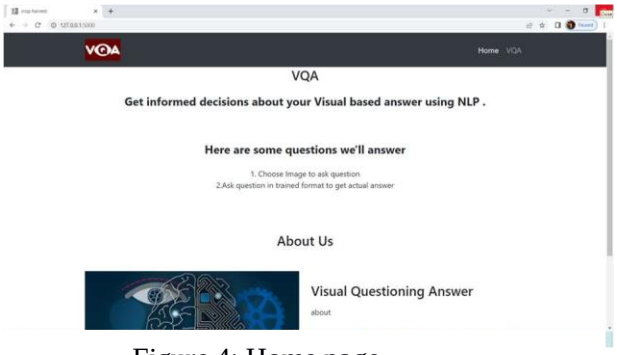
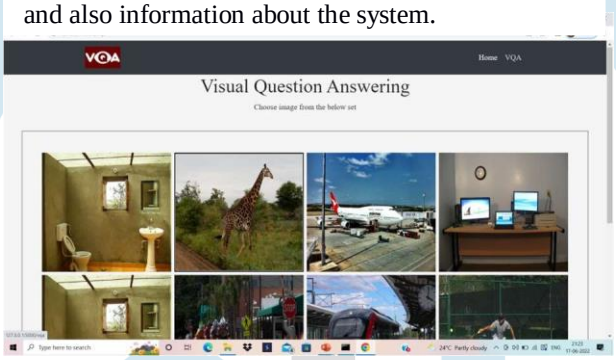


Figure 4: Home page

Figure 4 shows the snapshot of the home page which gives information about Visual Question Answering and also information about the system.



CONCLUSION

In this paper an experimental top-down and bottom-up collaborated visual question answering system is introduced. This visual attention mechanism enables the value to be calculated more naturally and with greater depth at the level of objects and other salient regions of the image. Putting this approach to the visual question answering system, have achieved futuristic and up to theminute results in our tasks, while simultaneously improving our capacity to grasp the required output. Thecorrectness of the answers is as expected and have got the results as desired. There can be a variety of questions that can ask and will get the appropriate answers accordingly. The pickle software, which greatly aided us in completing this job with such precision, also made handling the big data set easier.

REFERENCE

[1] K. P. Moholkar1, Ajay Pisharody2, Noorul Hasan Sayyed3, Rakesh Samanta4, Aadarsh Turkish: Visual Question Answering using Convolutional Neural Networks - Journal of Computer and Mathematics Education (TURCOMAT) Vol. 12No. 1S (2021)

[2] Yash Srivastava, Shiv Ram Dubey, nehasis Mukherjee and Vaishnav Murali: Visual question answering using deep learning: A survey and performance analysis. - International Conference on Computer Vision and Image Processing (2021)

[3] Data Sylvain Lobry, IEEE, Jesse Murray, Diego Marcos, Member, Devis Tuia, Senior Member: RSVQA- Visual Question Answering for Remote Sensing IEEE - IEEE Transactions on Geoscience and Remote Sensing, Volume: 58 (2020)Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nico- las Thome. Murel: Multimodal relational reasoning for vi- sual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Applying the method to VQA, establishing the wideapplicability of the approach. Recognition, pages 1989–1998 (2019)

[4] Sudan Jha, Vijender Kumar-Solanki, Raghvendra Kumar, Anirban Dey: A Novel Approach to VisualQuestion Answering Using Faster Region Based Convolutional Neural Networks for Parameter Prediction - International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 5, Nº 5 (2018)

[5] Yash Goyal, Douglas Summers-Stay, Tejas Khot, Devi Parikh and Dhruv Ba- tra. Making the v in vqamatter: Increasing the importance of picture comprehension in visual question answering. IEEEConference on Computer Vision and Pattern Recognition Proceedings, pages 6904– 6913 (2017)

[6] Ranjay Krishna, Oliver Groth, Yuke Zhu, Justin Johnson, Joshua Kravitz, Kenji Hata, StephanieChen, Li-Jia Li, Yannis Kalan- tidis, Li-Jia Li,Michael S. Bernstein, David A Shamma & Li Fei- Fei: Visual genome- Using crowdsourced dense picture annotations to connect language and vision - International Journal of Computer Vision volume123, pages32–73 (2017)

[7] Kevin J Shih, Derek Hoiem and Saurabh Singh. Where to look: Focus areas for answering visual questions. The IEEE conference on computer vision and pattern recognition published the results., pages 4613–4621 (2016)