IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Conceptual Framework For Visual Question And Answering System

[1]Siddesh Shimpi, [2]Himanshu Marathe, [3]Sarvesh Makane and [4]Dr. Sharvari S. Govilkar

[1,2,3]Student, [4]Head of Department
Department of Information Technology
Pillai College of Engineering, University of Mumbai, New Panvel, India

Abstract: Visual question answering is a field of study in the Artificial Intelligence domain. Visual Question Answering (VQA) is a trained system which provides an answer to the questions asked in respect to the given image in English Natural Language. VQA is a general system and has a power that can VQA be used in any image based scenario with enough training on the relevant dataset. To achieve this, neural networks are used, particularly Convolutional Neural Networks and Recurrent Neural Networks. In this research work, we have compared different approaches of VQA, out of which we are exploring CNN based models. This paper proposes the conceptual framework to a VQA model by using the combination of CNNs and RNNs under the computer vision tasks and NLP tasks.

Index Terms* - *VQA, Computer Vision , Natural Language Processing, Visual Feature, Textual Feature.

I.Introduction

Visual Question Answering (VQA) system is a model that takes an input as a single image and a single English natural language query about an image and generates an answer in the English natural language which is the output of the model. In this VQA model, the model has to perform reasoning over the contents of the image based on the query asked by the user. So, to answer if there are any specific animals, say lions in the image, the model must be able to detect animals and classify them into different types.Finally, to say which cricket player is batting, commonsense and reasoning and real world information are necessary. Taks like object recognition, object detection have been mentioned in the field of Computer Vision. A good Visual Question Answering system must be able to solve a broad range of Natural language processing and Computer Vision tasks, as well as thorough reasoning based on the image content.

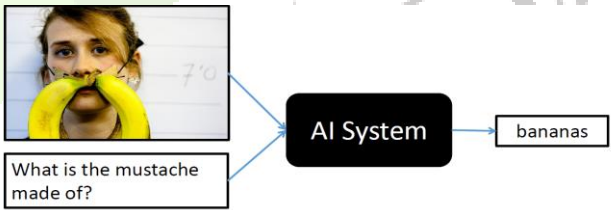

Fig 1.1 VQA

In the above figure, an input as an image of a girl wearing a mustache made up of banana is given, and when we ask a query asking what is the mustache made of, the model gives us the correct answer that it is made up of banana.

In this research paper, we have proposed a hybrid model approach consisting of (CV) computer vision tasks and (NLP) natural language processing tasks to extract the visual and textual features from the image and the question which will generate the accurate answer for the query.

II. Related Work

One of the works on visual question answering was started by Aishwarya Agrawal, Dhruv Batra, Devi Parikh, Aniruddha Kembhavi [1] which proposed a new system for VQA in which for every type of question, train and test sets have dissimilar distributions of solutions. They have presented two new datasets, VQA v1 and VQA v2, called Visual Question Answering under Changing Priors. Author evaluated many existing VQA models using the new system and presented that the performance reduces significantly compared to the original VQA system. Author finally proposed a Grounded Visual Question Answering model (GVQA) which contains restrictions and inductive biases in the architecture particularly developed to stop the model from 'cheating' by relying directly on the priors of the trained data. GVQA is made off an existing VQA model – Stacked Attention Networks (SAN). Their observations demonstrate that GVQA outstandingly outperforms SAN, a somewhat boost in the overall performance for both the VQA-CP v1 which gives accuracy of about 12.35% and VQA-CP v2 which gives accuracy of 6.34% on datasets. Author concludes that GVQA is greatly interpretable and transparent than existing VQA models and GVQA is an initial step towards developing models which are visually grounded by design. Future work concerns building models that can make the best use of both the worlds (visual grounding and priors), such as, giving answers to questions based on the knowledge about the world (sand is usually gold, plants are usually green).

Peter Anderson1, Chris Buehler, Xiaodong He, Damien Teney [2] have proposed a combined top-down and bottom-up attention mechanism that enables attention to be computed at the level of objects and other salient regions of the image. The bottom-up mechanism proposes image regions, each with a feature vector, while the top-down mechanism calculates feature weightings. In this paper the researchers have adopted the terminology where they refer to attention mechanisms driven by purely visual feed-forward attention mechanisms as 'bottom-up' and task-specific context as 'top-down'. The technique to implement a bottom-up attention model in "Bottom-Up Attention Model". In the VQA Model given a bunch of spatial features of the image V, our proposed VQA system model uses a 'soft' top-down attention mechanism to weight each feature, utilizing the question representation as context. The proposed model implements a familiar joint multimodal embedding of the image and the question, pursued by a prediction of regression of scores over a bunch of candidate answers. This paper combined the top-down and bottom-up visual attention mechanism. Applying this method to visual question answering and image captioning, they achieved modern results in both tasks, while improving the interpretability of the resulting weights.

Wenliang Cai, Guoyong Qui[3] have explained that this model uses the CNN and LSTM algorithms and the collaborative attention mechanism to generate a picture caption related to the problem information, and then combines the two text information on the image description and the question to obtain an answer and output the picture description. The results show that the proposed algorithm can predict answers more accurately. The visual question answer algorithm based on image caption proposed in this paper divides the image question answer system into two parts. Firstly, the image caption algorithm is used to obtain the image caption of the attention. In this paper the resulting picture caption is a description related to the problem, which is helpful for predicting the answer. Secondly, the obtained image caption is combined with the problem information, and then input them into the LSTM to predict the answer. This VQA-E Model first uses the image processing method in this model to extract the target information of the image, combine it with the text information, and use the collaboration attention mechanism in the process of combination, instead of the attention on image only in VQA-E model. Then the explanation is combined with the problem information and input into the LSTM system. The experimental results show that the visual question answering algorithm model based on image caption has good algorithm superiority. In further work, the research will adopt the migration learning method such as parameter fine-tuning to optimize the network to achieve better visual answer prediction.

K. P. Moholkar, Noorul Hasan, Ajay Pisharody, Aadarsh Valsange, Sayyed, Rakesh Samanta, [5] proposed a Deep learning model that is based on Convolutional Neural Networks (CNN) for the image feature and question answering tasks. The specified system is implemented in four parts :Extraction of features from image, creating question answer vocabulary, image module and prediction model. They have given the accuracy for different approaches used .Using knowledge based models they got accuracy of about 70%. They came to the conclusion that implementing the VQA model which is similar to human nature is a rewarding task.VQA is a very generalized system and hence it has a wide range of applications if specified datasets are available. Models based on knowledge base models and attention based models generate higher accuracy. They have also proposed a system which uses transfer learning which increases models ability to answer related questions.

Zeyuan Hu ,Jialin Wu and Raymond J. Mooney[6] addressed the issues of both Visual question answering and Image captioning and came up with a new model which compensated for both the issues. To solve these issues, they have presented a system which is efficient in jointly producing image captions and answering the visual questions. Hence they have implemented a model in which they make use of image and question to create question related captions and use them in the next step to provide knowledge to the VQA system. When they evaluated the process of adding captions as inputs in the vqa process, they observed huge improvements in accuracy over BUTD(Anderson et al., 2017). For image captioning tasks, they showed that the captions generated from their model show promising results as well as they provide more informative descriptions for vqa tasks.Their results on VQA v2 dataset accomplished 65.8% accuracy using generated captions and 69.1% accuracy using annotated captions When they evaluated their joint system against other state of the art methods, they found that their systems produce more informative captions and outperformed other state of the art systems in terms of VQA accuracy which indicate effectiveness of including captional features as additional inputs for the vqa system.

Luowei Zhou , Lei Zhang , Houdong Hu , Jason J. Corso1 , Jianfeng Gao Hamid Palangi[4] have proposed a unified Vision-Question-Answering(VQA) model. This model can be used for either image captioning tasks or visual question answering tasks which means it is unified in nature. ls. This unified model is pre-trained on a huge dataset consisting of numerous image-text pairs using the unsupervised learning. This is the first reported model which achieves results of higher accuracies on both vision-language generation tasks and language understanding tasks. Author was inspired by the success of pre-trained language models such as GPT AND BERT, both of which use a training scheme consisting of two phases. In their detailed study on VQA and Image Captioning, they have proved that use of unsupervised pre-training can massively improve a models accuracy by speeding up learning.

After the review of literature survey following gaps have been identified. From the review and survey papers, it has been identified that most of the research works are unable to comprehend the emotion of an agent. Many reviewed research works were not able to answer questions enquiring about the physical characteristics like size, age, or height of an object with a good accuracy. Few of the research works were not able to answer the questions which had multi-word answers accurately.

III. Visual Question And Answering system

User has to input an image file and input an English Natural Language Query. The image will pass through the Computer Vision(CV) task which will help to extract the visual features of the image. The query will pass through the NLP tasks which will extract the textual features of the query.

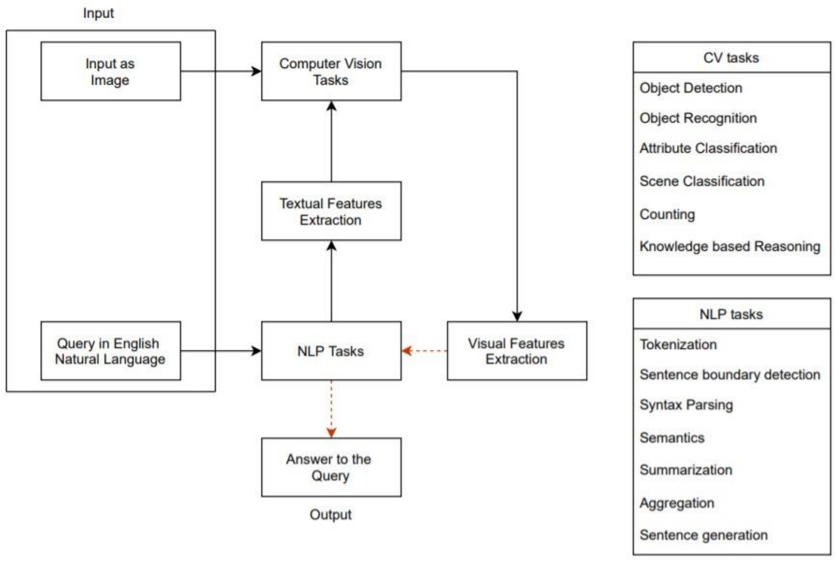

Fig. 1 Conceptual Framework for VQA System

1. **Computer Vision tasks:**

Computer vision is a field of study focused on the problem of helping computers to see. It is an integrative field that could in general be called a sub-field of machine learning and artificial intelligence, which involves the use of specialized techniques and make use of widely used learning algorithms.

1.1 Convolutional Neural Networks(CNNs): Convolutional Neural Network (CNN) is a sub-class of DNN which is generally applied to analyzing visual imagery. It is used not only in Computer Vision but also for text classification in Natural Language Processing (NLP). Most of the Computer Vision tasks are surrounded by CNN architectures, as the basis of most of the problems is to classify an image into known labels.

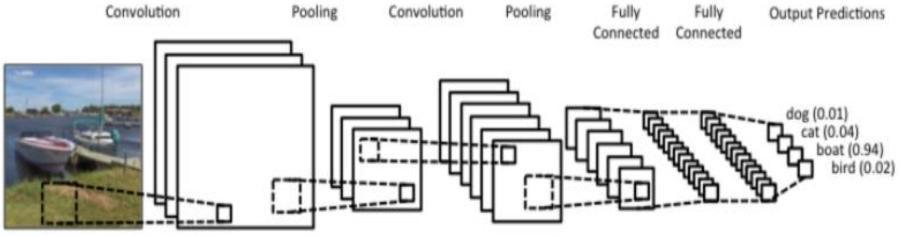

Fig. 2: CNN architecture

A CNN is a model used in machine learning to extract features, like texture and edges, from the image. What makes CNNs special is their ability to extract features from images. CNN is used for image classification and object segmentation.

1.2 Image Classification: Image classification is the process of calculating a particular class, or label, for something that is defined by a group of data points.

1.3 Object Detection: Object detection is the method of finding occurrences of real-world objects such as visual features, cars, and furniture in images. Object detection algorithms basically use extracted features and learning algorithms to identify occurrences of an object class.

1.4 Object Segmentation: It is the process of segregating an image into various regions based on the attributes of pixels to identify objects or boundaries to clarify an image and more efficiently predict it.

Object Recognition: Object recognition is a technique in which classification of objects taken from a digital image is performed.

1.5 Knowledge based Reasoning: Knowledge-based system (KBS) is a technique in AI that captures the features of human experts to provide logical decision-making abilities to a machine.

1.6 Recurrent Neural Networks(RNN): An RNN is a type of neural network that is designed to work with sequences. These sequences can be such as video, sound, text.

2. **Visual Feature Extraction:**

Visual Feature extraction is the main unit of the computer vision task. In fact, the complete deep learning model works about the idea of finding useful features which distinctly identifies the objects in the image. In computer vision, a visual feature is a piece of data in an image which is unique to the specific object. It may be a unique color in an image or a particular shape such as an edge, line, or an image region. A useful feature is used to differentiate objects from one another.

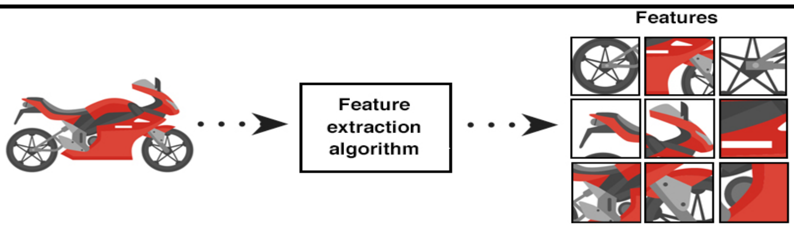

Fig. 3 Visual Feature Extraction

In the image that we have given above, we give the raw input image of a bike to a feature extraction algorithm. The extraction algorithm produces an array that contains a list of features of the images. This is called a feature vector which is a 1-Dimensional vector that makes a representation of the object given in the image.

3. Natural Language Processing tasks:

Natural Language Processing (NLP) is a specialized domain in artificial intelligence (AI). NLP assists a machine to understand human language and find out specific context which help them to perform tasks like machine translation, summarization etc automatically without any human assistance involved.

3.1 Tokenization: Tokenization is the splitting of a sentence or a textual document into smaller components, such as words or characters.These smaller components are called tokens.

3.2 Sentence Boundary Detection(SBD): Sentence boundary detection is a method of detecting where one sentence ends and another sentence begins.

3.3 Text Summarization: Text summarization is a technique which shortens a piece of text of which the main intention is to create an eloquent summary having only the main points contained in the document.

3.4 Semantics Analysis: Semantic analysis is a process of understanding human natural language and the context in which it is used.

3.5 Syntax parsing: Syntax parsing is a process of inspecting natural language with the rules of a formal natural language grammar.

4. Textual Feature Extraction:

Textual Feature extraction is a step in which information extraction from a textual context is done to represent a text message.Popular methods of textual feature extraction are:

4.1 Bag of words: Bag-of-Words is a technique used to transform a set of tokens into a set of features in reference to the context. The Bag of words module is used for document classification, in which each word is used as a factor for training the classifier.

4.2 TF-IDF: TF-IDF stands for term frequency-inverse document frequency. It is a technique in which highlighting of a word is done which has less frequency in textual documents but has a great importance. The TF–IDF value of a word increases proportionally to the frequency in which a word appears in the corpus.

5. Expected Output:

Output will be a word generated by the given proposed system. The features extracted by applying NLP tasks to the query will also be passed in Computer Vision tasks and Visual Feature Extraction will be done using the CV tasks which will generate the accurate answer for the given image and query.

IV. Conclusion

The conceptual framework of Visual Question Answering is discussed in detail in the aforementioned sections. Previous works related to VQA are studied thoroughly and analyzed for further improvements. Proposed architecture has two main tasks which include CV tasks and NLP tasks which extracts the visual and textual features from the image and the query respectively. The answer generated by the model with the help of these features has a good accuracy.

V. Acknowledgment

We would like to thank our Guide and Head of the Department, Dr Sharvari Govilkar for being ever supportive and helpful. Ma’am always pushed us to do better and showed immense faith in us. In the end we would like to thank our parents for pushing us to achieve our best and motivating us to strive for excellence. We couldn’t have gotten this far without their support.

REFERENCES

[1] A.Agrawal, D.Batra, D.Parikh and A.Kembhavi. 2018. Don’t Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. arXiv preprint arXiv:1712.00377v2.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. arXiv preprint arXiv:1707.07998.

[3] Wenliang Cai and Guoyong Qui. 2019. Visual Question Answering algorithm based on Image Caption. In Proceedings of the IEEE 3rd Information Technology,Networking,Electronic and Automation Control Conference (ITNEC 2019).

[4] Luowei Zhou, Hamid Palangi, Lei Zhang , Houdong Hu , Jason J.Corso and Jianfeng Gao. 2019. Unified Vision-Language Pre-Training for Image Captioning and VQA. arXiv preprint arXiv:1909.11059v3.

[5] K.P. Moholkar, Ajay Pisharody, Noorul Hasan Sayyed, Rakesh Samanta and Aadarsh Valsange. 2021. Visual Question Answering using Convolutional Neural Networks. In Turkish Journal of Computer and Mathematics Education Vol.12 No.1S (2021), 170-175.

[6] Jialin Wu, Zeyuan Hu and Raymond J. Mooney. 2018. Joint Image Captioning and Question Answering. arXiv preprint arXiv:1805.08389v1.