Topic -
Deep Learning Based Visual Question Answer(VQA)

Submitted By -:

Aman Soni

Abstract-

In order to answer fundamental 'common sense' questions on provided images, Visual Question Answer (VQA) tasks integrate data processing issues with languages and visual processing challenge. The VQA approach attempts to identify the correct answer to a question given a picture or a problem by combining visual elements of the image with information gleaned from textual queries. As a result, in this study, we analyse and cover current datasets released within the VQA sector that handle diverse question formats, the resilience of machine-learning models, and our VQA datasets have shown promise in terms of a deep learning model. Finally, some of the findings from our simple VQA model are shown and discussed.

In this report we will discuss about our project introduction and than define literature review after reading the paper . Than we discuss our problem statement and their solution statement and result and their conclusion and last we define some reference which are great help.

Introduction –

Visual Questions Answers is a field of study that focuses on creating an artificial intelligence system that can respond to questions posed in simple language on an image.

A system that completes this task demonstrates a deeper understanding of images: it must be able to reply to a wide range of questions about a picture, generally addressing different aspects of the image.

Using Visual Question Answering, the activities integrate data processing issues with language and visual processing to answer fundamental 'common sense' questions regarding given images .The VQA system uses visual elements of inference and visuals received through textual queries to try to find the right response to a question and an image in natural language, which is human language.

Literature Review –

- In the paper they initially covered significant datasets published for verifying the VQA in the study (such as visual7W,VQA etc.)
- Then Next they discussed about state of architecture designed for Visual Question Answer (such as Vanilla, Pythia v1.0 etc.)
- Then we computed result over Deep learning based VQA method Vanilla .
- In these paper they have used mainly 2 datasets:-

1. VQA Dataset
2. Visual7W Dataset

Problem Statement-

Looking at an image and answering any inquiry about it with our common sense understanding is straightforward for us. However,

there are situations when a visually impaired user or an intelligence analyst wants to deliberately elicit visual information from a photograph.

We have to create a deep learning system that can respond to open-ended questions about real-world photographs. For our tests, we employ the VQA dataset.To improve the findings, we have to investigate the usage of attention-based models.

Building a system/algorithm that accepts (ideally) any image and a question asked in natural language about that image and offers a natural language answer to that question with reference to the image as the output can be regarded as the broad problem of visual question answering (VQA).

Solution-

As a result, we develop a system that takes a picture as input and produces a natural language response in response to a free-form, open-ended, or natural language enquiry about the image.

Dataset:- The actual data set used to train the model to perform various actions is referred to as the data set**.** The types of datasets in VQA are DAQUAR ,Visual madlibs,CLEVER,Tally-QA,KVQA,Visual7W and VQA v1 and VQA v2.But in this report we use two datasets which is VQA(V1 or V2) and visual7W.

VQA contains two dataset VQA v1 and v2:-

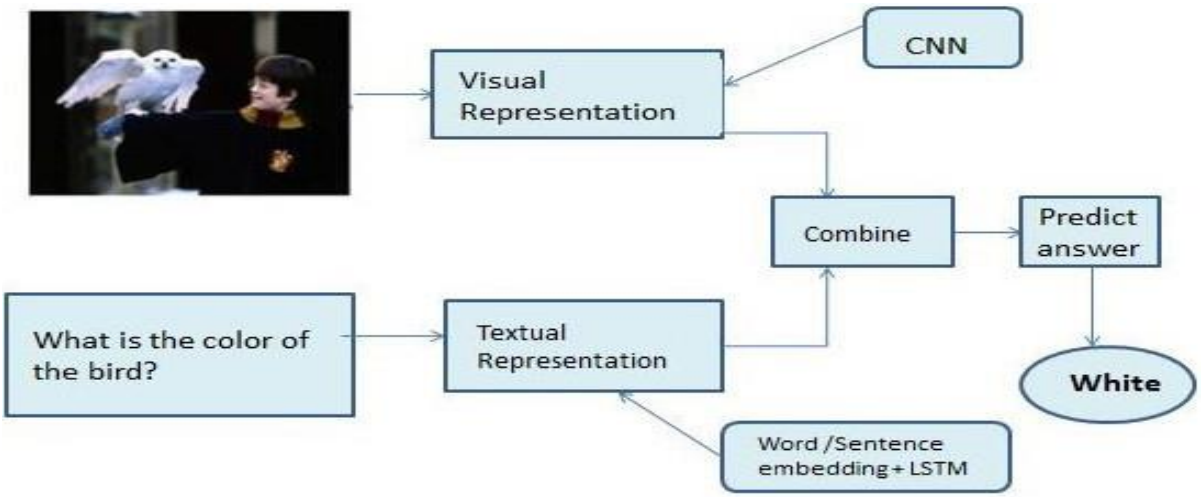
VQA v1 - It's one of the largest datasets, with photos from the COCO collection as well as some newly made images to complete the picture. There are three questions and about ten answers for each image. There are around 600 thousand QA-pairs, with over 200 thousand training and testing QA pairs and over 100 thousand validation QA pairs. There are roughly 82000 training images, 40000 validation images, and 81000 testing photos, as well as approximately 600 thousand QA-pairs.

VQAv2 - It was created to eliminate any biasness in the event that there is just one answer to a single question for a specific image. This was replaced with a different image that posed the same question but had a different response. As a result, the new dataset has doubled in size.

Visual7W - The MS-COCO dataset is also used in the Visual7W dataset . There are 327,939 question-answer pairs and 47,300 COCO images in it .There are also 1,311,756 multiple choice questions and answers in the collection, as well as 561,459 groundings.. What, Where, When, Who, Why, How, and Which are the seven sorts of questions addressed in the dataset (from which it gets its name). The majority of the questions fall into two groups. The text-based 'telling' questions provide a description.

Method -The method we are going to use is Vanilla VQA

Vanilla VQA : The vanilla VQA model, which is used as a foundation for deep learning methods, uses CNN to extract features and LSTM or RNN to interpret language. These characteristics are combined into a single feature that is used to categorise responses using element-wise operations.



Visual Question Answering necessitates image recognition and natural language process techniques; one major area of research is Deep Learning, which involves using Convolutional Neural Networks

4096 output units from last hidden layer 1024

(CNN) for image recognition, Recurrent Neural Networks (RNN) for combining the results of natural language processing with speech recognition to produce the final response, as shown in Figure.

The bounding boxes effectively indicated the group of possible responses.

Thes model's process is as follows:

CNN - CNN stands for Convolutional Neural Network and is a sort of artificial neural network used for image/object recognition and categorization. Using a CNN, Deep Learning recognises items in an image. CNNs are used for a number of activities and purposes, including image processing, computer vision tasks including localization and segmentation, video analysis, and obstacle recognition in self-driving cars, among others.

RNN –A recurrent neural network, or RNN, is a sort of artificial neural network that is used to process natural language and recognise voice. It's the most advanced algorithm for sequential data, and it's what Apple's Siri and Google's voice search employ. It is the first algorithm with an internal memory that remembers its input.

LSTM - Long short-term memory (LSTM) is a deep learning architecture that uses an artificial recurrent neural network (RNN).

LSTM has feedback connections, unlike traditional feedforward neural networks.It can handle not only single data points (such as photos), but also entire data streams (such as speech or video).

Result-

Consequently, The system will respond to a question in the following ways, as if it were a human:

1. From the inputs, it will learn visual and linguistic knowledge (image and question respectively)

2. bring the two data streams together

3. create the answer using this advanced knowledge

Conclusion –

We examine VQA using deep learning and research with a visually impaired user or an intelligence analyst wishing to actively elicit visual information from a picture. The most challenging difficulties in these areas that can be investigated using DL were also recognised in this study. According to the report, in the VQA or visual7W datasets, we attained an accuracy of around 55-60%. To achieve the same level of accuracy as the paper, we can train the model to be of higher quality and so achieve a higher degree of accuracy.

Reference –

- IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10,800–10,809, Synthesizing counterfactual samples for robust visual question answering (2020). H. Zhang, S. Pu, and Y. Zhuang. L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang.
- Empirical evaluation of gated recurrent neural networks for sequence modelling by J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. preprint arXiv:1412.3555 arXi (2014)
- S. Gupta, P. Arbelaez, and J. Malik: Perceptual organisation and interior scene detection from rgb-d images. IEEE CVPR, pp. 564–571. (2013)
- J. Schmidhuber and S. Hochreiter: Long-term memory. Neural Computation 9: 1735–1780 (8). (1997)