International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 07 | July - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Visual Question Answering Using Deep Learning

Mrs.Pallavi N R[1], Anusha A R[2], Chethan Kumar S[3], Hemanth G K[4], Reetha D[5]

[1] Mrs.Pallavi N R, *Asst. Professor, BGS Institute of Technology*
[2] *Anusha A R, Department of Computer Science and Engineering, BGS Institute of Technology*
[3] *Chethan Kumar S, Department of Computer Science and Engineering, BGS Institute of Technology*
[4] *Hemanth G K, Department of Computer Science and Engineering, BGS Institute of Technology ,*
[5] *Reetha D, Department of Computer Science and Engineering, BGS Institute of Technology*

Abstract – Answers to Practical Questions is a research center on how to build a computer program to answer questions asked in a clear and natural language. First, let us consider three key elements to answering the accompanying questions. Answering visual questions is a proposed task of integrating computer imaging and natural language processing (NLP), to facilitate research, and to push the boundaries of both fields. Computer programs, on the other hand, learn to detect, process, and process images. In short, their purpose is to teach the machines to see. NLP, on the other hand, is a field that allows for interaction between computers and people in a natural language, i.e. learning materials, among other activities. Both computer theory and NLP are the domain of artificial intelligence and share similar methods based on machine learning. Historically, however, they have developed separately. An effective way is to integrate convolutional neural networks (CNNs), which are trained in object recognition, and embedded words, trained in corporate writing.

***Key Words*:** *Natural Language Processing, Computer Vision, Convolutional Neural Network.*

1. INTRODUCTION

An answer to Visual Questions is a research site about building a computer program to answer questions posed in a graphic and natural language. First, let's examine the three databases in Responding to Visual Questions. Visual Question Answering (VQA) is a complete AI task set at the crossroads of computer vision (CV) and natural language processing (NLP). When looking at a pair of image questions, our model produces not only feedback, but also a set of reasons (such as text) and visual attention maps. VQA is a new database containing open-ended questions about images. These questions require an understanding of common sense, language and knowledge in order to be answered. Answering visual questions is a proposed activity to integrate computer vision and natural language processing (NLP), to facilitate research, and to push the boundaries of both fields. On the other hand, computer programming learns to find, process, and understand images. In short, their purpose is to teach machines to see. NLP, on the other hand, is a field that is about allowing interactions between computers and humans in a natural language, i.e. learning equipment, among other activities. Both computer theory and NLP are the domain of artificial intelligence and share similar methods based on machine learning. Historically, however, they have developed separately. Both sectors have seen significant improvements in their goals over the past few decades, and the combined growth of the visual and textual data explosion is leading to a marriage of effort from both sectors. An effective way is to integrate convolutional neural networks (CNNs), which are trained in object recognition, and embedding words, trained in corporate text.

Figure 1.VQA Model Diagram

The focus is on providing Visual Question Answering (VQA) a new skill model - the ability to read text on pictures and answer questions in consultation with text and other visual content. VQA has seen great progress. But today's VQA models fail miserably in questions that need to be studied. This is strange because these are the very questions that visually impaired users often ask on their apps. Specifically, a VizWiz study found that up to 21% of these questions involve reading and consulting text imagery of the user's environment - 'What is my oven setting?', „What is the Visual Question Answering (VQA), the computer is presented with an image and a textual question about this image. It must then determine the correct answer, typically a few words or a short phrase. Variants include binary (yes/no)] and multiple-choice settings, in which candidate answers are proposed. A closely related task is to “fill in the blank”, where an affirmation describing the image must be completed with one or several missing words. These affirmations essentially amount to questions phrased in declarative form. A major distinction between VQA and other tasks in computer vision is that the question to be answered is not determined until run time.

2. DATA SETS

DAQUAR: DAQUAR stands for Dataset for Question Answering on Real World Images, published by Malinowski et al. [20]. It was the first database released for IQA work. Images taken from NYU-Depth V2 database [27]. A small database with 1449 images. The interrogation bank includes a pair of 12468 questions with 2483 different questions. Questions are made up of personal annotations and enclosed within 9 question templates using NYU-In-depth data annotations.

VQA data set: The Visual Question Answering (VQA) data set [2] is one of the largest data sets collected in the MS-COCO database [18]. The VQA database contains at least 3 questions per image and 10 answers per question. The data set contains 614,163 queries in the form of open choice and more. For multiple choice questions, the answers can be divided into: 1) Correct Answer, 2) Reasonable Answer, 3) Popular Answers and 4) Random Answers. Recently, the VQA V2 [2] data set was released with additional confusing images.

Visual7W: The Visual7W database [39] is also based on the MS-COCO database. Contains 47,300 COCO images with 327,939 response pairs. The data list also contains 1,311,756 multiple choice questions and 561,459 based answers. The database deals mainly with seven types of questions (from which it gets its name): What, Where, When, Who, Why, How, and Where. It is made up of two types of questions. 'Telling' questions are those that are based on the scriptures, which provide a kind of meaning. 'Identifying' questions are the ones that start with 'Which,' and should be correctly marked with responsible boxes within the group of sound answers.

3. VQA-BASED-IN-DEPTH STUDY METHODS

Convolution Neural Network: The Convolutional neural network is a special type of server neural artificial feed network where communication between layers is facilitated by the visual cortex. Convolutional Neural Network (CNN) is a class of deep emotional networks used to analyze visual images. matrix and filter or kernel. Each input image will be transferred to a series of convolution layers with filters (kernels) to produce maps of the output feature. Here's how CNN works.

Basically, convolutional neural networks have 4 layers namely convolutional layer, ReLU layer, integration layer and fully integrated layer. Convolutional layer: It is in the convolution layer after the computer reads the pixel-shaped image, and with the help of the convolution layers we take a small piece of the image. These images or patches are called features or filters. By sending these solid features the same is almost the space in two images, the flexible layer is much better at seeing similarities than the scenes that accompany the image. These filters are compared with new input images when they are the same and the image is categorized accordingly. Here list the features and image and duplicate each image, pixel by matching pixel element, add pixels at the top and divide the total number of pixels in the element. We create a map and place the filter values in the corresponding area. Similarly, we will move the feature to all other areas of the image and see how the feature matches that location. Eventually, we will get the matrix as output.

ReLU Layout

The ReLU layer is nothing but a fixed line unit, in this layer it removes all negative value from filtered images and replaces zero. This is done to prevent the prices from being reduced to zero. This is a conversion function that only makes a node if the input value exceeds a certain number while the input is less than zero the output will be zero and subtract all negative values from the matrix.

Blend layer

In this layer we reduce or decrease the image size. Here we first select the window size, then specify the required step, and then navigate through your window in your filtered images. Then take the maximum values in each window. This will cover the layers and reduce the image size and matrix. A reduced size matrix is provided as an input to a fully connected layer.

Fully Connected Layer

We need to compile all the layers after transferring it to the convolutional layer, ReLU layer and integration layer. Fully connected layer used for image capture. These layers need to be repeated if necessary unless you get a 2x2 matrix. Then finally a fully integrated layer is applied when the actual separation occurs.

4.RESULTS OF ASSESSMENT AND REVIEW

The reported effects of different methods on different databases are summarized in Table 1 and Table 2. It can be noted that the VQA data set is often used in different ways to test performance. Other data sets like Visual7W, Tally-QA and KVQA are also very challenging and the latest data sets. It can also be seen that Pythia v1.0 is one of the latest and most effective methods on the VQA database. Differential Network is the most recently proposed version of VQA work and shows promising performance across a variety of databases.

As part of this survey, we also used different methods for different databases and conducted experiments. We have considered the following three models for our evaluation, 1) the Vanilla VQA basic model [2] using the VGG16 CNN ar- Visual Question Answering using In-Depth Learning: A Survey and Performance Analysis.

The test results are presented in Table 3 with the accuracy of three models in two databases. In experiments, we found that Teney et al. [31] is a highly efficient model in both the VQA and Visual7W Dataset. The], most recently, where they used the same model with additional layers to improve performance. The accuracy of VQA is very low due to the nature of the problem. VQA is one of the most difficult computer vision problems, where the network has to understand the semantics of images, questions and relationships in the feature. Advanced Design is one of the most important stages of software development. The purpose of the design phase is to plan the solution to the problem specified by the required document. This phase is the first step from the root of the problem to the root of the solution. System design is probably the most critical aspect affecting software quality and has a huge impact on the latest phases, especially testing and maintenance. The result of this section is design text. The design work is usually divided into two distinct categories, system structure and detailed design.

A system design, sometimes called a top-level design, aims to identify the modules that should be in the system, the details of

these modules and how they work together to produce the desired result. At the end of the system design, all major data structures, file formats, output formats and major modules in the system and its specification are determined. High-level design provides a review of the entire system, identifying all its features to a certain degree of inaccessibility.

The design consideration briefs about how the system behaves for the boundary environments and what action should be taken if the abnormal case happens. Some of the design considerations are data collection, pre-processing methods, classification and association. Dataset is used to select an image. To extract the features of image using CNN algorithm. By using RNN & LSTM algorithm given question is processed. Outputs of CNN & RNN are merged by attention mechanism to produce an answer. Several recent papers have begun to study the answers to the practical questions. However, unlike our work, these are appropriately limited (sometimes artificial) settings with minimal databases. For example, it only considers questions whose answers come from the pre-defined closed world of 16 basic colors or 894 categories of objects. And considers questions generated in illustrations from consistent vocabulary of objects, features, interrelationships, etc. In contrast, our proposed work involves open, frivolous questions and responses. Our goal is to maximize the variety of information and thinking needed to provide the right answers. Essential to achieving success in this difficult and unforgettable task, our VQA data set is two orders larger than - (> 250,000 vs. 2,591 and 1,449 images respectively) . The proposed VQA project has a link to another related task: read the shared video analysis and related text to answer questions on two databases containing 15 video clips each. Uses multi- resource staff to answer questions about visual content asked by visually impaired users. In the same work, it is proposed to integrate the LSTM query with CNN so that the image can produce the answer - the same model is tested in this paper. Create invisible to assess the feasibility of a general concept vaccine. Presented a data set of 10k images and ordered captions describing specific aspects of the scene (e.g., individual items, what will happen next). At the same time as our work, questions and answers were collected in Chinese (later translated into English) for COCO photographs. Automatically generates four types of queries (object, number, color, location) using COCO captions. Text-based Q&A is a well-studied problem in NLP and text-based societies (recent examples that Other related text functions include sentence completion (e.g., with multiple choice answers). These methods provide inspiration for VQA strategies. for example, integrated text descriptions and QA pairs based on imitating characters and objects in a consistent environment (images). Our questions are man- made, which makes the need for common knowledge and complex thinking even more important. VQA related activities are tagging, photo captions and video captions, where words or sentences are produced to describe the visual content. Although these tasks require visual knowledge and semantic knowledge, captions are often not. The questions in VQA require precise information about the image. Other Vision and Language Functions. We now define the data set for Visual Question Answering (VQA). We begin by describing the actual pictures and the invisible scenes used to collect the questions. Next, we describe our process of collecting questions and related answers.Assessment of questions and answers collected asand the basic results are given in the following sections. An Real Pictures. We use 123,287 training and certification 81,434 photographs and test images from the recently released Microsoft Common Objects in Context (MS COCO).

5. CONCLUSION

The Visual Question Answering recently proved to be of great interest and advancement to a team of researchers and scientists from around the world. The latest trends seen in the development of data sets look at real life more and more by combining the type of questions and answers of the real world. Recent trends are also evident in the development of in-depth complex learning models by making better use of visual and textual features in a variety of ways. The performance of the best model remains and is only about 60-70%. Therefore, it remains an open challenge to develop better in-depth learning models and challenging data sets for VQA. Different techniques like object level details, separation mask, in-depth models, question feelings, etc. can be considered for the development of next-generation VQA models. And the total number of frames is also validated. The process endswhen the current frame reaches the last frame. Otherwise, itwill continue to detect anomalous activity. Hj calculates how much the feature affected the map that was tracking thefeature vector. The extracted feature is the influencedensity of the motion influence map.

Advanced Design is one of the most important stages of software development. The purpose of the design phase is to plan the solution to the problem specified by the required document. This phase is the first step from the root of the problem to the root of the solution. System design is probably the most critical aspect affecting software quality and has a huge impact on the latest categories, system structure and detailed design. A system design, sometimes called a top-level design, aims to identify the modules that should be in the system, the details of these modules and how they work together to produce the desired result. At the end of the system design, all major data structures, file formats, output formats and major modules in the system and its specification are determined. High-level design provides a review of the entire system, identifying all its features to a certain degree of inaccessibility. This is in contrast to the low-level design that reveals the detailed structure of these features.

6. REFERENCE

[1] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., Vol. 2. Oxford: Clarendon, 1892, pp. 68-73.

[2] I. S. Jacobs and C. III, G. T. Rado and H. Suhl, Eds. New York: Education, 1963, pages 271-350.

[3] K. Elissa, "The title of the paper if known," which has not been published.

[4] R. Nicole, “The title of a paper with only the first letter,” J. Name Stand. Abbrev., By pressing.

[5] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electronic spectroscopy studies in magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp.

740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[6] M. Young, Technical Writer. Mill Valley, CA: University Science, 1989

[7] Karpathy and L. Fei-Fei. In-depth Synchronization of Semantic Visualization of Image Definitions. In CVPR, 2015.1

[8] J.-H. Kim, S.-W. Lee, D.-H. Kwak, M.-O. Heo, J. Kim, J.- W. Ha, and B.-T. Zhang. Multimodal Residual Learning for Visual QA. In NIPS, 2016. 2

[9] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Integrating Visual Embedding-Semantic's With Neural Multimodal Language Models. TACL, 2015. 1

[10] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.- J. Li, D. A. Shamma, et al. Visual genome: Links language to visual cues using duplicate images from crowd sources. arXiv preprint arXiv: 1602.07332, 2016. 2