International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 3, pp: 42-47 www.ijaem.net ISSN: 2395-5252

Visual Question Answering

Manjusha K, Binu R
Computer Science and Engineering GEC Palakkad Malappuram, India
Computer Science and Engineering GEC Palakkad Palakkad, India
Corresponding Author: Manjusha K

Date of Submission: 15-07-2020 Date of Acceptance: 31-07-2020

ABSTRACT—Visual Question Answering is a system that takes an image and natural language question about the image as an input and generates natural language answer as an output. Visual Question Answering (VQA) is the hot topic in natural language processing Computer Vision and Deep Learning. Visual question answering means if given an image and a natural language question related to that image then the model will predict the answer in natural language. The challenge of visual question answering is that different questions require different types and levels of understanding of an image to find correct answers. Here examine a Visual Question Answering System with the combination of Image Captioning and Question Answering. This project contains mainly two parts Image Captioning and Question Answering. Here uses a deep learning approach with convolutional neural network(CNN) and Recurrent neural network(RNN) that is LTSM to implement the Image Captioning part and BLEU score helps for the evaluation of image captioning and Flickr8k dataset is used. And for Question Answering uses an open sourced new technique for NLP called BERT with SQuAD dataset. The output of image captioning is used as an input for question answering. When combining the result of these two will get the Visual Question Answering model. The model will provide an image caption and can ask multiple questions based on the caption then the model will predict the answers in natural language.

Index Terms—CNN, LSTM, RNN, BERT

I. INTRODUCTION

Visual Question Answering (VQA) [1] is a system that takes an image and natural language question about the image as an input and generates natural language answer as an output. Given an image with the corresponding question then it must able to understand the image well in order to generate an appropriate answer. Suppose the question is based on the number of items then the system must be able to detect the number of items and for answering the color then the system needs to detect the color. It has been solved in the field of Computer Vision with good result. The common approach of combining convolutional and recurrent neural networks is to map images and questions to a common feature space. The Visual Question Answering model uses a classical CNN-LSTM model where image features and language features are computed separately and combined together and a multi-layer perceptron is trained on the combined features. The VQA implementation field is so complex because it needs a good dataset like VisDial, DAQUAR, COCO-QA, and VQA-Dataset. Here examine a Visual Question Answering System with the combination of Image Captioning and Question Answering.

This project contains mainly two parts Image Captioning and Question Answering. Here uses a deep learning approach with convolutional neural network (CNN) and Recurrent neural network (RNN) that is LTSM to implement the Image Captioning part and BLEU helps for the evaluation. And for Question Answering uses an open sourced new technique for NLP called BERT [2] with SQuAD dataset. When combining the result of these two will get the Visual Question Answering model with high speed. In this era VQA is the main topic and so many research works and VQA challenge conferences are happening. The Image Captioning and Question answering will give huge boost to the Visual Question Answering field.

II. RELATED WORK

The image captioning problem and its proposed solutions have existed since the advent of the Internet and have huge boost in this era. Its widespread adoption as a medium to share images. Numerous algorithms and techniques have been put forward by researchers from different perspectives for both image captioning and question answering. Krizhevsky et al. [5] implemented a neural network using convolutional neural and a very efficient unique method GPU implementation of the

convolution function. By employing a regularization method called dropout, they succeeded in reducing over fitting problem. Their neural network consisted of multiple layers convolutional layer, maxpooling layers, ReLu and a fully connected layer. The final 1000-way softmax as the decoder. Deng et al. [6] introduced a new database which they called ImageNet, an extensive collection of images that contain many images built using the core of the WordNet structure.

ImageNet organized the different classes of images in a densely populated semantic hierarchy and it will be helpful for better result. Karpathy and FeiFei [7] made use of datasets of images contain many images and their sentence descriptions to learn about the inner correspondences visual data and language from the images. The work described a Multimodal Recurrent Neural Network architecture that uses long short term memory and utilizes the inferred co-linear arrangement of features in order to learn how to generate novel descriptions of images. Yang et al. [8] proposed a system for the automatic generation of a natural language description of an image, which will help immensely in furthering image understanding like object detection, color identification, pattern recognition, number of items etc. The proposed multimodal neural network method,

Consisting of object detection and localization modules, is very similar to the human visual system which is able to learn how to describe the content of images automatically from the image for captioning. Then the description is used as an input for question answering. In order to address the LSTM units being complex and inherently sequential across time and which is then fed into the decoder side, Aneja et al. [9] proposed a convolutional network model which contain multiple layers for feature extraction for machine translation and conditional image generation and also helps for feature extraction.

Pan et al. [10] experimented extensively with multiple net- work architectures like CNN, RNN on large datasets consisting of varying content styles, captions, and proposed a unique model showing improvement on captioning accuracy over the previously proposed models and also uses the bleu score for the evaluation. Vinyals et al.[11] presented a generative model consisting of a deep recurrent architecture that contains long short term memory and that leverages machine translation and computer vision, used to generate natural descriptions of an image based on the extracted features by ensuring highest probability of the generated sentence to accurately describe the target image.

Xu et al. [12] introduced an attention based model that learned to describe the image regions automatically, i.e. extracted features from that image. The model was trained using standard back propagation techniques by maximizing a variable lower bound and it will avoid over fitting. The model was able to automatically learn identify object boundaries while at the same time generate an accurate descriptive sentence and will provide image caption for that image. The system proposed Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol the task of free-form and open-ended Visual Question Answering (VQA). Given an image and a natural language question about the image, then the model will provide the answer in natural language. Mirroring real-world scenarios, such as helping the visually impaired, both the questions and answers are open-ended they can understand the environment around them. Visual questions selectively target different areas of an image including background details like colors, object detection, number of items and underlying context. As a result, a system that succeeds at VQA typically needs a more detailed understanding of the image with extracted features and complex reasoning than a system producing generic image captions. Based on the caption question answering will works. Moreover, VQA is amenable to automatic evaluation, since many open-ended answers contain only a few words or a closed set of answers that can be provided in a multiple-choice format and provide more accuracy.

Abhishek Das, SatwikKottur, Khushi Gupta, Avi Singh [13] proposed the task of Visual Dialog, which requires an AI agent to hold a meaningful dialog with humans in natural, conversational language about visual content like a coversational agent. Specifically, given an image, a dialog history, and a question about the image, then the model will give the corresponding answer when the agent has to ground the question in image, infer context from history.

The question accurately based on the extracted features and understanding of the image. Visual Dialog is disentangled enough from a specific downstream task that should be give the corresponding answers like a chatbot so as to serve as a general test of machine intelligence, while being grounded in vision enough to allow objective evaluation of individual responses and benchmark progress it will provide more accuracy. The model will develop a novel two-person chat data-collection protocol like a chatbot, where can ask multiple questions related with the image then the model should give the answer in natural language

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 3, pp: 42-47 www.ijaem.net ISSN: 2395-5252

to curate a large-scale Visual Dialog dataset (VisDial). The model is a conversational chatbot for visual question answering.

III. DATASET

A. Flickr8k

In the past few years, the problem of generating descriptive sentences automatically for images has garnered a rising interest in natural language processing deep learning and computer vision . Image captioning is a fundamental task which requires semantic understanding of images like colors, object detection, number of items etc and the ability of generating description sentences with proper and correct structure. It will showcase the efficiency of proposed model using the Flickr8K datasets and show that the model gives superior results compared with the state-of-the-art models utilizing the Bleu metric. Flickr 8k dataset contain more than 8000 images and each image is annotated with 5 captions in the dataset. Where 6000 images are used for training and 1000 images each used for testing and validation for better result. Flickr 8K is a dataset consisting of 8,092 images from the Flickr.com website with captions. This dataset contains collection of day-to-day activity with their 5related captions for each image.

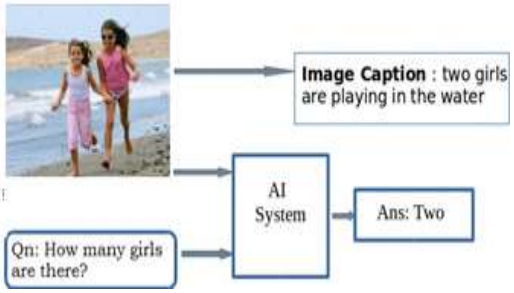
B. SQuAD

SQuAD is the most used and latest question answering dataset and is a reading comprehension dataset consisting of questions posed by crowd workers on a set of Wikipedia articles. And the answers to every question are a segment of text or span from the corresponding reading passage or reference texts. SQuAD combines the 100,000 questions over 50,000 new unanswerable questions which is written adversarial by crowd workers to look similar to answerable ones and will give a better result.

IV. METHODOLOGY

The Visual Question Answering system is based on Deep learning approaches like CNN and an LSTM. Question Answering (QA) systems have emerged as powerful platforms for automatically answering questions asked by humans in natural language and also for image captioning. Visual question answering is a challenging task that has received increasing attention from the computer vision, deep learning and the natural language processing communities. The critical challenge of this problem is that different questions require different types and levels of understanding of an image.

And find correct and accurate answers. Visual Question Answering means if we given an image and a natural language question related to that image then the model will predict the answer in natural language. If an image is given then the model will predict an image caption, then the obtained caption is used as an input for the question answering system. Then can ask multiple questions based on that caption then question answering system will give the corresponding answer in natural language. Here implementing a visual question answering that contains two part, an image captioning part and a question answering part. In this project uses a deep learning approach with convolutional neural network (CNN) and an LTSM to implement the Image Captioning part and BLEU score helps for the evaluation. And for Question Answering uses an open sourced new technique for NLP called BERT with SQuAD dataset. When combining the result of these two will get the Visual Question Answering model.



A. Image Captioning

The Image Captioning can be done by using Convolutional neural network (CNN) and an LSTM (Long short term memory) followed by softmax. The image features will be extracted from CNN model, trained on Flickr dataset that contain more than 8000 images and 5 captions for each image and then feed the features into the LSTM model which will be responsible for generating the image captions for corresponding image. CNN are specialized deep neural networks which can process the data that has input shape like matrix and have huge boost in this era. There are four basic building blocks in CNN. Convolution layer, ReLu layer, pooling layer and fully connected layer and after that will get a vectorized form of features. CNN is used to extract the image features from the image. Long short term memory is the type of RNN which is suited for sequence prediction problems, i.e. based on the previous text, can predict what the next word will be based on the probability.

CNN model extract the features from the image and LSTM translate the features and objects given by the image based sentence in text form and

International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 3, pp: 42-47 www.ijaem.net ISSN: 2395-5252

will provide a better result. In the past few years, the problem of generating descriptive sentences automatically for images has garnered a rising interest in natural language processing, deep learning and computer vision research. Image captioning is a fundamental task which requires semantic understanding of images.

Generating description sentences with proper and correct structure and have huge boost in this era. In this method the use of multi layer Convolutional Neural Network (CNN) to generate vocabulary describing the images and a Long Short Term Memory (LSTM) to accurately structure meaningful sentences using the generated keywords. The convolutional neural network contain four layers and it compares the target image to a large dataset of training images, then generates an accurate description using the trained captions. It showcases the efficiency of our proposed model using the Flickr8K datasets and show that their model gives superior results compared with the state-of-the-art models utilizing the Bleu score. The Bleu metric is an algorithm for evaluating the performance of a machine translation system by grading the quality of text translated from one natural language to another for better performance prediction. The performance of the proposed model is evaluated using standard evaluation matrices. BLEU score which outperform previous benchmark models.

- Feature Extractor: The CNN is used to extract the features and the feature extracted from the image has a size of 2048 with a dense layer. It will be reduce the dimensions to 256 nodes in each layer. The feature of the images from the Flickr 8K dataset is extracted using the CNN model due to the performance of the model in object identification. The convolutional neural network which consists of 16 layer which has a pattern of 2 convolution layers followed by 1 dropout layers until the fully connected layer at the end. The dropout layers are present to reduce overfitting the training dataset for more accuracy, as this model configuration learns very fast. These are processed by a dense layer to produce a 4096 vector element representation of the photo and passed on to the LSTM layer and based on the probability it will predict the next word.
- Sequence Processor: An embedding layer will handle the textual input followed by the LSTM layer, identifying the context of the words and represented in the form of real value vectors. LSTM layer to process the text data which is now in vector form after it pass through embedding layer in order to preserve the sequence information of text and find out the correlation among the different word features and predict the next word based on the probability. The function of a sequence processor is for handling the text input by acting as a word embedding layer in the sequence processor. The embedded layer consists of rules to extract the required features of the text and consists of a mask to ignore padded values also. The network is then connected to a LSTM for the final phase of the image captioning and gets the correct answer.
- Decoder: By merging the output from the above two layers it will then process by the dense layer to make the final prediction or to predict the captions based on the image. The final phase of the model combines the input from the Image extractor phase or extracted features and the sequence processor phase using an additional operation then fed to a 256 neuron layer and then to a final output Dense layer that produces a softmax, i.e the decoder prediction of the next word in the caption over the entire vocabulary which was formed from the text data that was processed in the sequence processor phase and based on the probability of words it will predict the next word. Then finally get the image caption for the corresponding image.

B. Question Answering

Question Answering uses an open sourced new technique for NLP called BERT and it is a major breakthrough which took the deep learning community by storm because of its incredible performance of BERT. BERT which stands for Bidirectional Encoder Representations from transformers and it is can be used in classification word sense disambiguation, summarization etc. It is a pretrained deep learning representation model. Using pretrained BERT models can utilize pretrained memory information of sentence structure, language and text grammar related memory of large corpus of millions or billions of annotated training examples that it has trained and also it will be the latest trending in question answering system. In the pretraining step, There is a fixed vocabulary and a tokenizer. Where the text is split into tokens and mapped to their index in the tokenizer vocabulary.

At the end of every sentence needs to append special [SEP] token and map the tokens to their ids. And also POS tagging and word embedding The pretrained model can then be fine-tuned on small data tasks like Question Answering,

classification, sentiment analysis, entity recognition, word sense disambiguation etc. BERT is a huge model with 24 transformer blocks, 1024 hidden units in each layer and 340M parameters and it is the latest method for question answering. The model is pretrained on 40 epochs over a 3.3 billion word corpus including Book corpus (800 million words) and English Wikipedia (2.5 billion words) and have huge boost in deep learning. BERT model is well defined in understanding the given text summary and answering the question from that summary and here image captioning is the text summary. To understand the question related information BERT has trained on SQuAD dataset. Both the question and the reference text are the input for this BERT mode.

The question and reference text or images captioning text are separated by [SEP] token. BERT uses segment embeddings to differentiate the question from reference text. This is then added to token embedding before feeding into input layer. BERT needs to highlight a span of text that containing the answer. This is represented as simply predicting which token marks the start of the answer and which token marks the end for finding the result. For every token in the text, feed its final embedding into the start token classifier. The start token classifier only has a single set of weights represented by blue rectangle which it applies to every word. After taking the dot product between the output embeddings and the start weights, apply the softmax activation to produce a probability distribution over all of the words.Whichever word has the highest probability of being the start token is the one that will pick. Repeat the process for end token then will get separate weight vector.

V. RESULT AND DISCUSSION

In order to evaluate the image-caption pairs, need to evaluate their ability to associate previously unseen images and captions with each other. BLEU score check how close the generated text or caption is to expected text or caption using sentence bleu from nltk. The evaluation of model that generates natural language sentence can be done by the BLEU (Bilingual Evaluation Understudy) Score and inside he initialization there is two list first on is for actual description and other for predicted description. It is widely used to evaluate performance of Machine translation. Sentences are com- pared based on modified n-gram precision method for generating BLEU score and finds the bleu score for 1 gram, 2 gram, 3 gram and 4 gram. Here got a bleu score .5 and .6. It is a better result.

BERT is bilingual evaluation understudy score and is conceptually simple and empirically powerful. For question answering the BERT model will get a mean F1 score of 95.9 with precision 98.4 and recall 98.2. Where the precision and recall are metrics that have been accepted by the research community of search effectiveness. In the context of Question answering system the precision is the proportion of retrieved answers that are correct and recall is the proportion of correct answers that are retrieved. F1 score is an accuracy measure of precision and recall combined. It measures the average overlap between the prediction and the correct answer. So the model will get a best result.

VI. CONCLUSION

Visual question answering contains an image captioning part and a question answering part. Image captioning is a deep learning approach for the captioning of images. The Keras was used with Tensorflow as a backend to implement the deep learning architecture to achieve better result and have an effective BLEU score for the model. The Bilingual Evaluation Understudy Score is a metric for evaluating a generated sentence to a reference sentence. In the future the working on alternating Pretrained Photo Models to improve the feature extraction of the model. Also, there is a planning to improve better performance by using word vectors on a much larger corpus of data such as news articles and other online sources of data. The configuration of the model was tuned, but other alternate configurations can be trained to see for improvement in the performance of the image captioning model. The question answering is done by using BERT which is the major breakthrough in this era. Once gets the result of image captioning then the result is used as an input for question answering model. Based on the caption then can ask multiple questions the model will predict the answer in natural language.

REFERENCES

[1]. J. Wu, Z. Hu, and R. J. Mooney, "Joint image captioning and question answer- ing," ArXiv, vol. abs/1805.08389, 2018.

[2]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv: 1810.04805, 2018.

[3]. H. Wang, Y. Zhang, and X. Yu, "An overview of image caption generation meth- ods," Computational Intelligence and Neuroscience, vol. 2020, 2020.

[4]. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh,

"Vqa: Visual question answering," in Proceedings of the IEEE inter- national conference on computer vision, pp. 2425–2433, 2015.

[5]. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Infor- mation Processing Systems 25 (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012.

[6]. J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255, 2009.

[7]. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3128–3137, 2015.

[8]. D. S. Lakshminarasimhan Srinivasan and A. Amutha, "Image captioning-a deep learning approach," International Journal of Applied Engineering Research, vol. 13, no. 9, pp. 7239–7242, 2018. G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[9]. J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in Proceedings of the IEEE conference on computer vision and pattern recogni- tion, pp. 5561–5570, 2018.

[10]. J.-Y. Pan, H.-J. Yang, P. Duygulu, and C. Faloutsos, "Automatic image captioning," in 2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763), vol. 3, pp. 1987–1990, IEEE, 2004.

[11]. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156–3164, 2015.

[12]. C. Liu, J. Mao, F. Sha, and A. Yuille, "Attention correctness in neural image captioning," in Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[13]. A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, "Visual dialog," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 326–335, 2017.

[14]. Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image caption- ing and visual question answering based on attributes and external knowledge," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 6, pp. 1367–1381, 2017.

[15]. A. Gulli and S. Pal, Deep learning with Keras. Packt Publishing Ltd, 2017.

[16]. R. Shanmugamani, Deep Learning for Computer Vision: Expert tech- niques to train advanced neural networks using TensorFlow and Keras. Packt Publishing Ltd, 2018.

[17]. J. Deka, Image Captioning: Capsule Network vs CNN approach. PhD thesis, Dublin, National College of Ireland, 2020.

[18]. G.-L. Chao and I. Lane, "Bert-dst: Scalable end-to-end dialogue state track- ing with bidirectional encoder representations from transformer," arXiv preprint arXiv: 1907.03040, 2019.