Visual Question Answering using Convolutional and Recurrent Neural Networks

Ankush Azade[1], Renuka Saini[2], and Dinesh Naik[3]

National Institute of Technology Karnataka, Surathkal, India-575025
[1]azadeankush8@gmail.com,[2]renukasaini.202it021@nitk.edu.in,[3]din_nk@nitk.edu.in

Abstract. This paper presents a methodology that deals with the task of generating answers corresponding to the respective questions which are based on the input images in the dataset. The model proposed in this methodology constitutes two major components and then integration of analysis results and features from these components to form a combination in order to predict the answers. We have created a pipeline that first preprocess the dataset and then encodes the question string and answer string. Using NLP techniques like tokenization and stemming, text data is processed to form a vocabulary set. A yet another experiment with modification in model and approach was performed using easy-VQA dataset which is available publically. This model used bag of words technique to turn a question into a vector. This approch considered two component separately for text and image feature extraction and merging it to form analysis and generate answer. Merge is done by using element-wise multiplication. In these approaches we have used softmax activation function in the output layer to generate output or answer to the question. When compared to existing methodologies this approach seems comaparable and gives decent results.

Keywords: CNN, Question-Answering, VQA

1 Introduction

"Visual Question Answering" is a topic that inculcates the input as an image and a set of questions corresponding to a particular image which when fed to neural networks and machine learning models generate a answer or multiple answers. The purpose of building such systems is to assist the advanced tasks of computer vision like object detection and automatic answering by machine learning models when receiving the data in the form of images or in even advanced versions, receiving as video data. This task is very essential when we consider research objectives in Artificial intelligence. In recent developments of AI[1], the importance of image data and integration of tasks involving textual and image form of input is huge. Visual question answering task will sometimes we used to answer open ended questions, otherwise multiple choice or close ended answers. In our methodology we have considered the formulation of open ended answers instead of close ended because in real world, we see that most of the human interactions involve non binary answers to questions. Open ended questions are a part of much bigger pool of the set of answers, when compared to close ended, binary or even multiple choice answers.

Some of the major challenges that VQA tasks face is computational costs, execution time, and the integration of neural networks for textual and image data. It is practically unachievable and inefficient to implement a neural network that takes into account both text features and image features and learn the weights of the network to make decisions and predictions. For the purposes of our research we have considered the state-of-the art dataset which is publically available. The question set that could be formed using that dataset is very wide. For instance one of the questions for an image containing multiple 3-D shapes of different color can be "How many objects of cylinder shape are present?"[1]. As we can see this question pertains a very deep observation, similar to human observation. After observing, ,experimenting and examining the dataset questions we could see that each answer requires multiple queries to converge to an answer. Performing this task requires knowledge and application of natural language processing technquies in order to analyse the textual question and form answers. In this paper, we discuss about the model constructed using Convolutional Neural Network layers for processing image features and Recurrent Neural Network based model for analysing text features.

2 Literature Survey

A general idea was to take features from a global feature vector by convolution network and to basically extract and encode the questions using a "lstm" or long short term memory networks. These are then combined to make out a consolidated result. This gives us great answers but it fails to give accurate results when the answers or questions are dependent on the specific focused regions of images.

We also came across use of stacked attention networks for VQA by Zichao Yang(2016)[3] which used extraction of semantics of a question to look for the parts and areas of the picture that related to the answer. These networks are advanced versions of the "attention mechanism" that were applied in other problem domains like in image caption generation and machine translation,etc. The paper by Zichao[3] proposed a multiple layer stacked attention network.
This majorly constituted of following components : (1) A model dedicated to image, (2) a separate model dedicated to question, which can be implemented using a convolution network or a Long Short Term Memory (LSTM) [8] to make out the semantic vector for questions, and (3) the stacked attention model to see and recognise the focus and important part and areas of the image. But despite of its promising results, this approach had its own limitations.

A research by Kexin Yi et al.[4] in 2018 proposed with a new model, this model had multiple parts or components to deal with images and question/answers. They made use of a "scene parser", a "question parser" and something to execute. In the first component, Mask R-CNN was used to create segmented portions of the image. In the second component meant for the question they used a ”seq2seq” model. The component used for program execution was made using modules of python which would deal with the logical aspects of the questions in the dataset.

Focal visual-text attention for visual question answering (Junwei Liang et al., 2019)[5] This model (Focal Visual Text Attention) combines the sequence of image features generated by the network, text features of the image and the question. Focal Visual Text Attention used a hierarchical approach to dynamically choose the modalities and snippets in the sequential data to focus on in order to answer the question, and so can not only forecast the proper answers but also identify the correct supporting arguments to enable people validate the system's results. Implemented on a smaller dataset and not tested against more standard datasets.

Visual Reasoning Dialogs with Structural and Partial Observations (Song-Chun Zhu et al., 2019) [7] Nodes in this Graph Neural Network model represent dialogue entities (title, question & response pairs, and the unobserved questioned answer) (embeddings). The edges reflect semantic relationships between nodes. They created an EM-style inference technique to estimate latent linkages between nodes and missing values for unobserved nodes. (The M-step calculates the edge weights, whereas the E-step uses neural message passing (embeddings) to update all hidden node states.)

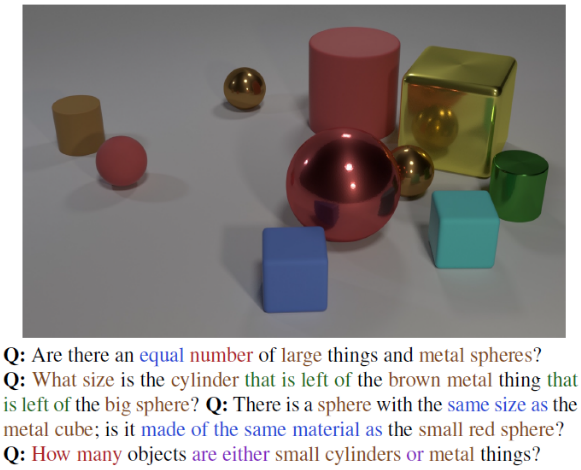
3 Dataset Description

The CLEVR10(”A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning”)[2] dataset was used, which includes a 70,000-image training set with 699,989 questions, a 15,000-image validation set with 149,991 questions, a 15,000-image test set with 14,988 questions, and responses to all train and val questions. Refer Dataset-1 statistics from Table 1 and a sample image from Fig. 1.

Table 1. Dataset-1 Statistics

	Train	Validation	Test
Image	70,000	15,000	15,000
Question	699,989	149,991	14,988

For Experminet-2 we have used a dataset titled easy-vqa which is publically available. This dataset is a simpler version of the CLEVR dataset, it mainly contains 2-Dimensional images of different shapes with different colors and positions. Dataset Statistics can be referred from Table 2 and a sample image from the easy-vqa dataset from Fig.2.



Q: Are there an equal number of large things and metal spheres? **Q:** What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere? **Q:** How many objects are either small cylinders or metal things?

Fig. 1. Sample Image from Dataset-1

Table 2. Dataset-2 Statistics

	Train	Test
Image	4,000	1,000
Question	38,575	9,673
Binary Questions	28,407	7,136

4 Proposed Method

After reading about multiple techniques and models used to approach VQA task, we have used CNN+LSTM as the base approach for the model and work our way up. 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 questions are encoded using a two-layer LSTM, while the visuals are encoded using the last hidden layer of CNN. After that, the picture features are l2 normalised. Then the question and image features are converted to a comman space and we have taken a element wise multiplication to obtain a answer. As a part of another approach we have used CNN based model architecture for image feature extraction and for text features extraction bag of words technique has been used to form a fixed length vector and simple feed forward network to extract the features. Refer Fig. 3 for proposed model.

4.1 Experiment 1

CNN A CNN takes into account the parts and aspects of an input fed to the network as an image. The importance termed as weights and biases in neural

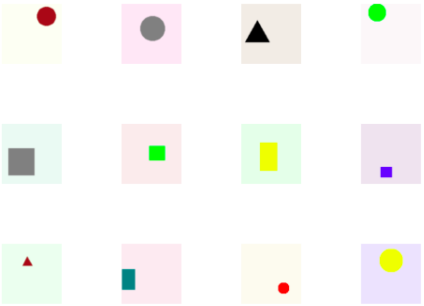


Fig. 2. Sample Image from Dataset-2

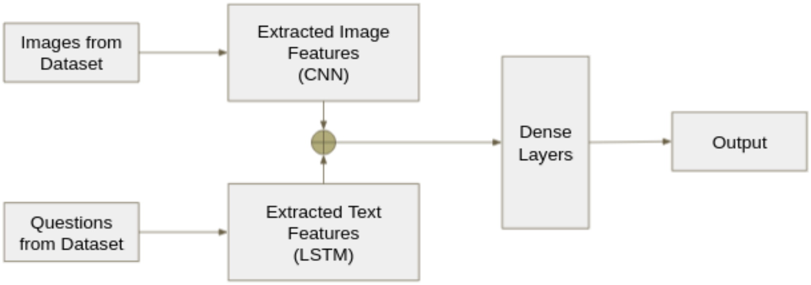

Images from Dataset
Extracted Image Features (CNN)
Questions from Dataset
Extracted Text Features (LSTM)
Dense Layers
Output

Fig. 3. Proposed Model

networks is assigned based on the relevance of the aspects of the image and also points out what distinguishes them. A ConvNet requires far less pre-processing than other classification algorithms. CNN model shown in Fig. 4 . We have used mobilenetv2 in our cnn model. MobileNetV2 is a convolutional neural network design which as the name suggests is portable and in other words ”mobile-friendly”. It is built on an inverted residual structure, with residual connections between bottleneck levels. MobileNetV2[9] is a powerful feature extractor for detecting and segmenting objects. The CNN model consists of image input layer, mobilenetv2 layer and global average pooling layer.

MobileNetV2 In MobileNetV2, there are two types of blocks. A one-stride residual block is one of them. A two-stride block is another option for downsizing. Both sorts of blocks have three levels. 1x1 convolution using ReLU6 is the initial layer, followed by depthwise convolution. The third layer employs a 1x1 convolution with no non-linearity.

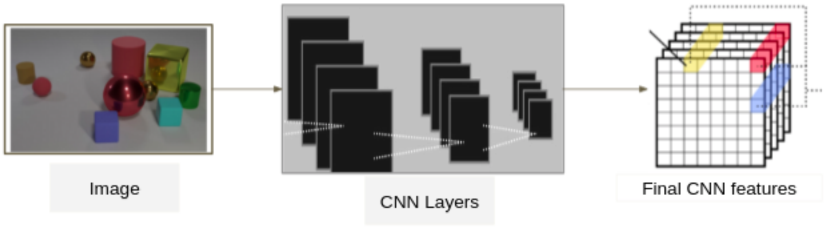


Fig. 4. Convolutional neural network

LSTM In sequence prediction problems, LSTM networks are a type of recurrent neural network that can learn order dependency. Given time lags of varying lengths, LSTM is ideally suited to identifying, analysing, and forecasting time series. The model is trained via back-propagation. Refer Fig.5. LSTM model consists of text input layer, one embedding layer and three bidirectional layers consisting of LSTM layers.

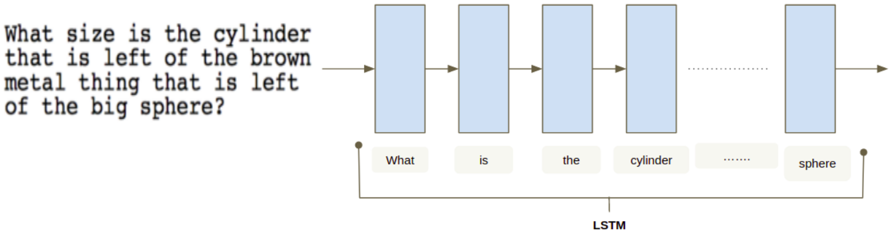

What size is the cylinder that is left of the brown metal thing that is left of the big sphere?
What
is
the
cylinder
....
sphere
LSTM

Fig. 5. Recurrent Neural Network

After implementation of CNN and the LSTM model we take their outputs and concatenate them.

$$Out = Multiply([x1, x2]) \tag{1}$$

where,
x1 = Output from CNN,
x2 = Output from LSTM,
Out = Concatenation of x1 and x2

After this we will create a dense layer consisting of softmax activation function with the help of tensorflow. Then we will give CNN output, LSTM output

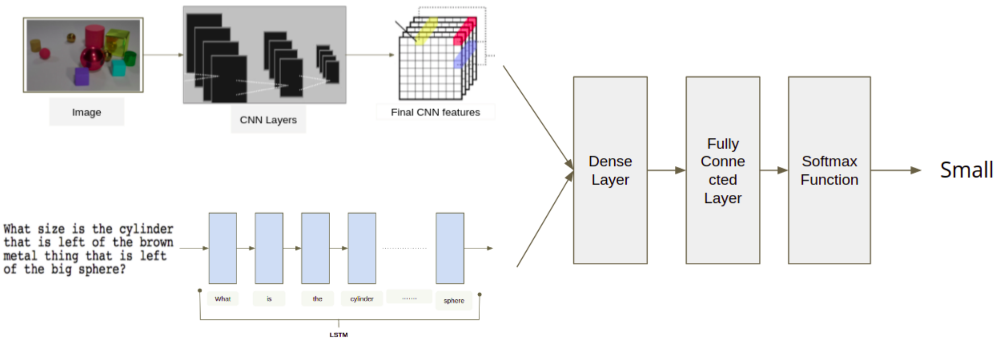


Fig. 6. Visual Question Answering

and the concatenated dense layer to the model. Refer Fig.6 for overall architecture. The adam optimizer and sparse categorical cross entropy loss were used to create this model. For merging the two components we have used element wise multiplication and fed it to the network to predict answers.

4.2 Experiment 2

As a first step we have preprocessed both the image data and the text data i.e. the questions given as input. For this experiment we have used a CNN model for extracting features from image dataset. In the fig.8, we have represented the model architecture used in the form of block representation. The input image of 64*64 is given as the input shape and fed to further layers. Then through a convolution layer with eight 3x3 filters using "same" padding, the output of this layer results in 64x64x8 dimension. Then we used a maxpooling layer to reduce it to 32x32x8, further the next convolution layer uses 16 filters and generates in 32x32x16. Again with the use of maxpooling layer, it cuts the dimension down to 16x16x16. And finally we flatten it to obtain the output of the 64x64 image in form of 4096 nodes. Refer Fig.7.

In this experiment instead of using a complex RNN architecture to extract the features from the text part that is the questions. We have used bag of words technique to form a fixed length vector and simple feed forward network to extract the features refer 8. The figure below represents the process. Here, we have passed the bag of words to two fully connected layers and applied tanh activation function to obtain the output. Both these components have been merged using the elementwise multiplication as discussed in previous section as well.

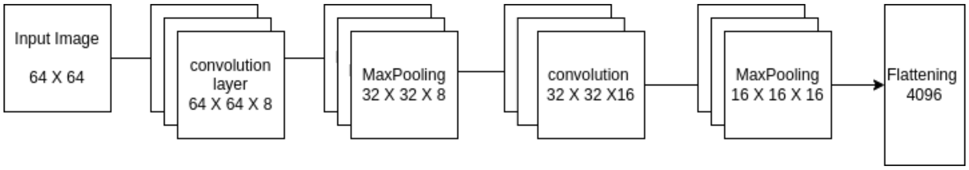


Fig. 7. CNN - Experiment 2

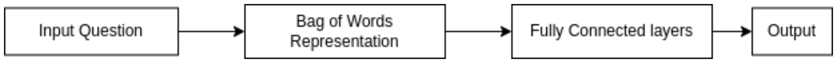


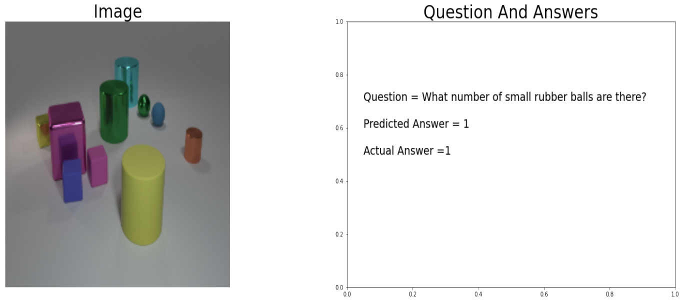
Fig. 8. Text Feature Extraction- Experiment 2

5 Results and Analysis

Following are the results for the experiment 1 and experiment 2.

5.1 Experiment 1:

From the below Fig.9 and Fig.10 we can see that, in the given image there are few solid and rubber shapes having different colors. For this respective image we have a question "What number of small rubber balls are there". For this question we have actual answer as 1. and our model also predicts the value as 1 which is correct.



Question = What number of small rubber balls are there?

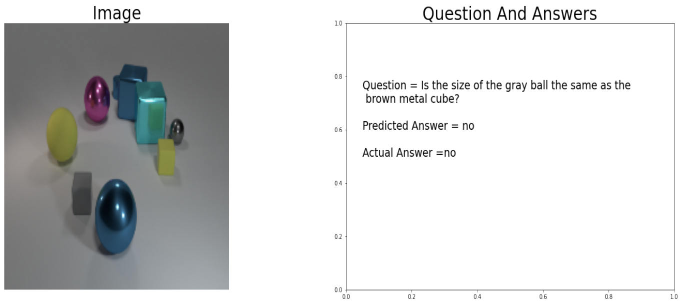
Predicted Answer = 1

Actual Answer =1

Fig. 9. Results of Experiment 1 (a)

5.2 Experiment 2:

In the second experiment we have considered a simpler form of the CLEVR dataset. And as explained in the methodology uses different model and variation



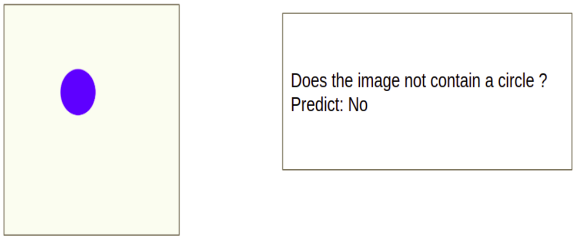
Question = Is the size of the gray ball the same as the brown metal cube?

Predicted Answer = no

Actual Answer =no

Fig. 10. Results of Experiment 1 (b)

of the approach. In the below Fig.11 we can see that the we have given a image and for that image we have a question ”Does this image not contain a circle ?” and our model predicted the correct answer as ”No”.



Does the image not contain a circle ?
Predict: No

Fig. 11. Results of Experiment 2

Observing the gradual increase in accuracy with each epoch with positive changes shows us that there is learning happening in our model at each step. Since calculating the accuracy for a vqa task is not objective because of open ended nature of the questions. We have achieved a training accuracy of 90.01% and test accuracy of 85.5% 3, this is a decent result when compared to the existing methodologies[1]. These results were observed on easy-VQA dataset.

Table 3. Train and Test Accuracy

Epoch	Train Accuracy	Test Accuracy
1	67.79%	72.69%
2	74.68%	76.89%
3	76.55	77.20%
4	77.77	77.87%
5	79.10	79.09%
6	82.17	81.82%
7	85.28	83.32%
8	87.02	83.60
9	88.40%	84.23%
10	90.01%	85.5%

6 Conclusion

Visual question answering result analysis is a subjective task. We used two component approaches which after performing separate extractions, merged its findings to obtain a consolidated result and predict the open ended answers. It can be concluded that the approach performed well and that the use of CNN network is very essential for image feature extraction. And also the use natural language processing techniques is essential for question feature extraction. Compared to baseline models the strategy is similar with tweaks discussed in the methodology section proved to be working well for a visual question answering system.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., Parikh, D. (2015). Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision (pp. 2425-2433).
2. Dataset: https://visualqa.org/download.html
3. Yang, Z., He, X., Gao, J., Deng, L., Smola, A. (2016). Stacked attention networks for image question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 21-29).
4. Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., Tenenbaum, J. (2018). Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. Advances in neural information processing systems, 31.
5. Liang, J., Jiang, L., Cao, L., Li, L. J., Hauptmann, A. G. (2018). Focal visual-text attention for visual question answering, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6135-6143).
6. Wu, C., Liu, J., Wang, X., Li, R. (2019). Differential Networks for Visual Question Answering. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 8997-9004. https://doi.org/10.1609/aaai.v33i01.33018997
7. Zheng, Z., Wang, W., Qi, S., Zhu, S. C. (2019). Reasoning visual dialogs with structural and partial observations, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6669-6678)
8. https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/

9. https://towardsdatascience.com/review-mobilenetv2-light-weight-model-image-classification-8febb490e61c
10. Yun Liu, Xiaoming Zhang, Feiran Huang, Xianghong Tang, Zhoujun Li, Visual question answering via Attention-based syntactic structure tree-LSTM, Applied Soft Computing, Volume 82, 2019, 105584, https://doi.org/10.1016/j.asoc.2019.105584, (https://www.sciencedirect.com/science/article/pii/S1568494619303643)
11. Nisar, Riddhi, Bhuva, Devangi, Chawan, Pramila. (2019). Visual Question Answering using combination of LSTM and CNN: A Survey, 2395-0056
12. Chen, Kan, Wang, Jiang, Chen, Liang-Chieh, Gao, Haoyuan, Xu, Wei, Nevatia, Ram. (2015). ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering
13. Neha Sharma, Vibhor Jain, Anju Mishra.: An Analysis Of Convolutional Neural Networks For Image Classification, Procedia Computer Science, Volume 132, 2018, Pages 377-384, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.05.198, (https://www.sciencedirect.com/science/article/pii/S1877050918309335)
14. Staudemeyer, R. C., Morris, E. R. (2019).: Understanding LSTM–a tutorial into long short-term memory recurrent neural networks, arXiv:1909.09586.
15. Md. Zabirul Islam, Md. Milon Islam, Amanullah Asraf, A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images, Informatics in Medicine Unlocked, Volume 20, 2020, 100412, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2020.100412. (https://www.sciencedirect.com/science/article/pii/S2352914820305621)
16. Wadii Boulila, Hamza Ghandorh, Mehshan Ahmed Khan, Fawad Ahmed, Jawad Ahmad.: A novel CNN-LSTM-based approach to predict urban expansion. Ecological Informatics, Volume 64, 2021, https://doi.org/10.1016/j.ecoinf.2021.101325, (https://www.sciencedirect.com/science/article/pii/S1574954121001163)