



Visual Question Answering

Dissertation submitted in partial fulfilment for the award of the degree

Master of Technology in Computer Science

by

TARUN BORANA

Roll No.: CS2017

M.Tech, 2nd year

Under the supervision of

Dr. Ujjwal Bhattacharya

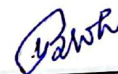
Computer Vision and Pattern Recognition Unit

INDIAN STATISTICAL INSTITUTE

July, 2022

CERTIFICATE

This is to certify that the work presented in this dissertation titled "Visual Question Answering", submitted by Tarun Borana, having the roll number CS2017, has been carried out under my supervision in partial fulfilment for the award of the degree of Master of Technology in Computer Science during the session 2021-22 in the Computer Vision and Pattern Recognition Unit, Indian Statistical Institute.



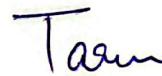
Dr. Ujjwal Bhattacharya , MSc MPhil PGDCA PhD
Associate Professor, Computer Vision and Pattern Recognition Unit
Indian Statistical Institute, Kolkata

Acknowledgements

First and foremost, I take this opportunity to express my sincere thankfulness and deep regard to *Dr. Ujjwal Bhattacharya*, for the impeccable guidance, nurturing and constant encouragement that he had provided me during my post-graduate studies. Words seem insufficient to utter my gratitude to him for his supervision in my dissertation work. Working under him was an extremely knowledgeable experience for a young researcher like me.

I shall forever remain indebted to my parents, teachers and friends for supporting me at every stage of my life. It is their constant encouragement and support that has helped me throughout my academic career and especially during the research work carried out in the last one year.

Date: 05-07-2022



Tarun Borana
Roll No.: CS2017
M.Tech, 2nd year
Indian Statistical Institute

Abstract

In recent years, tremendous progress has been made in the fields of object detection, computer vision, and natural language processing. Artificial intelligence Systems (AI), such as question-answering models provide the machine with "comprehensive" capabilities using natural language processing. Such a machine can respond to queries in natural language about an unstructured text. For performing the task of VQA, we can combine Natural language processing with computer vision. The purpose of a visual question answering system is to create a system capable of answering natural language queries about images. A number of systems have been introduced for visual question answering that use learning algorithms and deep-learning architectures.

This project introduces a VQA system that uses deep understanding of images using a deep convolutional neural network (CNN) that helps to extract features from image and LSTM are used for word embeddings for question texts. In this project we are taking only those questions that have answer type yes or no. Hence, Our system achieves complex reasoning and natural language understanding so that it can correctly predict the request and give the appropriate answer yes or no. Different architectures are introduced to combine the image and language models.

Contents

1	Introduction	3
2	Problem definition	5
3	Terminology	7
3.1	Neural Network Architectures	7
3.2	Feature Engineering	9
3.3	Evaluation Metrics	10
4	Existing Techniques and Related Work	11
4.1	Datasets	11
4.2	Feature Engineering	11
4.3	Long short-term memory (LSTM):	12
4.4	Related Work	12
5	Proposed Method for Visual Question Answering	14
6	Results, Conclusion, and Future Work	18

List of Figures

1	Sample Image	5
2	High-level view of a VQA system	6
3	Example of simple neural network with two hidden layer	7
4	Example of a simple RNN	8
5	Original image	8
6	Convolution filter	8
7	An example of CNN with 1 input layer,2 hidden layer,1 output layer,3 convolution layer, 2 max pooling layer, 4 features map in hidden layer 1, 8 features map in layer 2 and 4 different classes	9
8	Histogram between No. of words in ans and no. of ans(e.g 1)Is there any humans in the image ? ans:Yes,2)Which sport is being played ? ans: Table tennis	14
9	Histogram between no. of words in question and no of question , Minimum and maximum length of question is 5 and 17 words	15
10	Model Architecture of visual question answering with answer type yes and no	16
11	Model architecture for CNN with LSTM	17
12	Accuracy vs epoch for model1	18
13	Loss vs epoch for model1	19
14	Accuracy vs epoch for model2	19
15	Loss vs epoch for model2	20
16	Accuracy vs epoch for model3	20
17	model3	21
18	Some examples(1)Is there furniture in the room? Predicted Answer:No,2) Are both people on the bench? Predicted Answer:Yes)	21
19	Some examples(1)Is the desk cluttered? Predicted Answer:No,2)Are there clouds in the picture? Predicted Answer:Yes)	22

List of Tables

1	Accuracy and Loss	18
---	-------------------	----

1 Introduction

Artificial intelligence (AI) technology has a wide range of applications and has been used extensively to build parts of bigger systems since the 1990s. The advancement of deep learning has led to an exponential development in AI applications. In essence, deep learning is a branch of AI. AI can be used for a wide variety of activities, including object recognition, computer vision, machine translation, and understanding.

Large-scale research is now being done to address issues that connect several branches of AI. These projects could be categorised as multidisciplinary. "Image caption" might be a basic illustration of this. It combines basic computer vision (CV) algorithms to determine the broad idea of the image with natural language processing methods, like n-grams to create a caption for this picture. While being a multidisciplinary AI task, image captioning, The AI challenge doesn't require the system to fully understand the natural language or the internal workings of the image in order to generate the captions.

One of the tasks that require a very advanced and deep understanding natural language is the task of answering a question (QA). There is a concept under computer science, machine learning, and deep learning using natural language processing (NLP), text analysis, and information retrieval to create a system that can answer questions in natural language. Responding to a request for any reading material (e.g. contents from articles, blogs, passages, etc.) requires modeling complex interactions between context and request. The mechanism of the question-answering system is similar to the way humans understand, and therefore it is known as "machine comprehension". . In order to get a detailed understanding of the semantics of the reference text and query, we use the current state-of-the-art question answering system using advanced deep learning architectures and then use various methods to capture the relation between reference text and query. Sometimes it is also important that the model understands about general knowledge or common sense so that it could answer a query correctly.

In recent years, visual answers to questions (VQA) is another area of research that has generated a lot of interest in the era of artificial intelligence. VQA can be seen as an extension of the concept of machine understanding. It is also interdisciplinary AI a task that combines advanced computer vision and NLP to build a system capable of responding to a request for an image. A model tries to understand the meaning and semantics of the image and answer questions based on its understanding.

Unlike image captions, where the main knowledge of CV and natural language processing is enough to build AI models, tasks such as VQA needs a thorough knowledge of the best practices in both of these areas. Until now, most of the AI systems created could not be compared to humans at a high level due to the lack of ability for deeper reasoning. But now it is possible to try to build a system that should excel at tasks like VQA with the ongoing research in this field. Such a system usually requires a more detailed understanding of the image and sophisticated reasoning abilities. A more advanced version of this system will also have factual knowledge to be able to understand and answer questions that are not directly answered in the image.

VQA is a relatively new and interesting concept. Most extensive research happening in this area uses various deep learning and learning architectures algorithms in CV, object recognition, and NLP to achieve the goal. Most existing VQA datasets and models built on them tend to focus on questions that can be answered by direct analysis of input images. If we observe the general problem, we see that tasks such as VQA represent a very diverse set of problems and challenges to be overcome under the auspices of vision, language, and semantic representation of knowledge. We can build different components of VQA systems with existing research in these individual areas and then integrate them to solve the problem.

This project aims to build such a visual question-answering system. In this report, we formulate the problem with subsequent discussion of existing models, related research papers , and existing datasets available to VQA. We then provide an overview of what we offer systems, experiments, and their results.

2 Problem definition

Building an artificial intelligence system that can answer a question about a given Image is one of the most popular areas of research in recent years. Let's Consider the image shown below:



Figure 1: Sample Image

Ideally, the VQA system should be able to answer questions about the given image. Some examples of questions :

- Which sports are they playing ?
- Is there any human in the picture ?
- How many boys are there ?

Most of the questions like above are answerable by people without any major difficulties. A question like "How many boys are there ?". It is not the hard task for humans to count the books and give the answer "3". But for an AI system, it is extremely difficult to understand the meaning of the question and understand the semantics of the image, identify the relationship between request and image, and then try to respond to Question.

However, with the advent of research in deep learning, both in areas CV and NLP, we can now create such a system that understands and is able to answer these questions

correctly and with excellent results.

In general, one can identify the general problem of (VQA) is how to build a system/algorithm that accepts (ideally) any image and a natural language query about the given image and gives a natural language answer to this a question as an output. General view of such a system shown in figure 2.

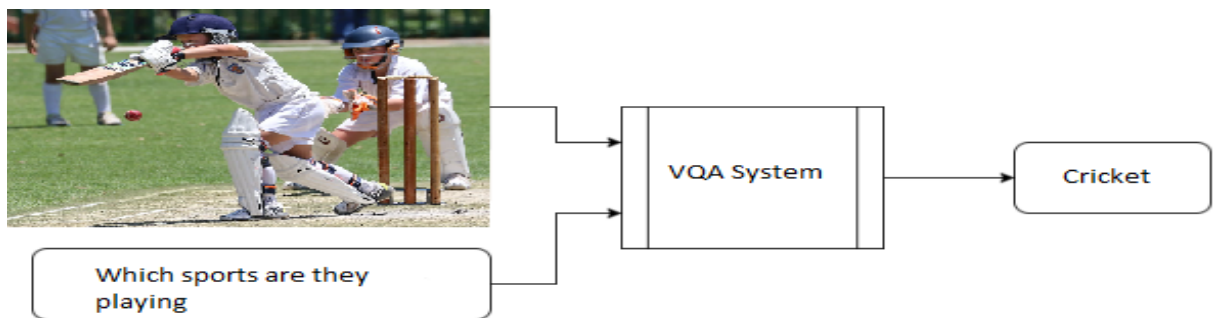


Figure 2: High-level view of a VQA system

3 Terminology

3.1 Neural Network Architectures

An artificial neural network (ANN) is a network consisting of many artificial neurons connected to each other in accordance with specific network architecture. The task of a neural network is to transform the input data into meaningful output data. The ANN architecture consists of one input layer, one output layer and a series hidden layers. See figure 3.

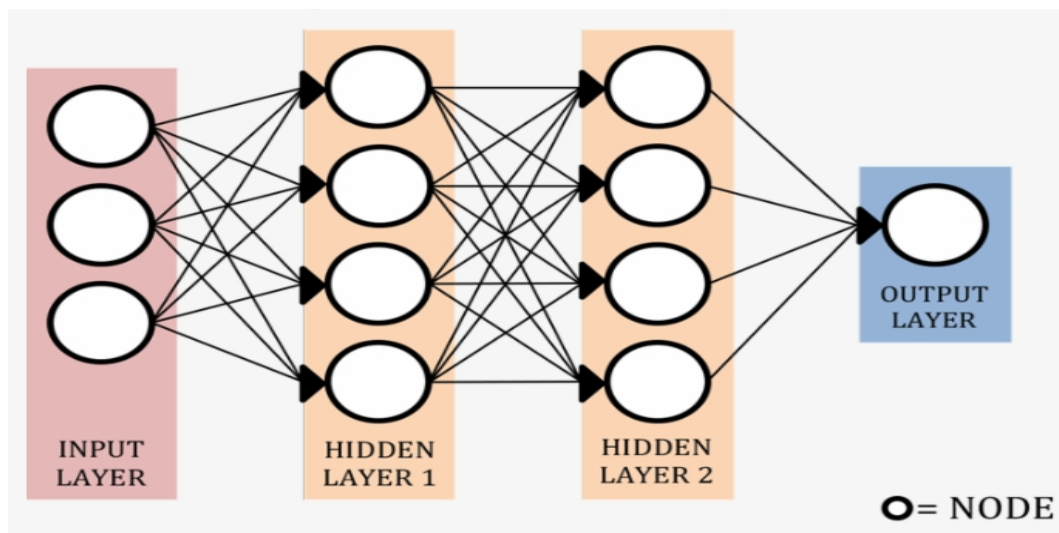


Figure 3: Example of simple neural network with two hidden layer

- **Recurrent Neural Network:** Recurrent Neural Networks (RNNs) are those that perform the same task for each element of the input sequence. The RNN output depends on the current element and previous computation. RNNs have "memory" to capture all the information they had seen before. A simple RNN is shown in Figure 4
- **Convolution:** Convolution is a mathematical operation that is performed on two functions and that are taken as input. It usually gives the integral of the dotted product of two input functions. Intuitively, it can be viewed as a modified version of one of the original functions in relation to another function. Consider a 5x5 matrix (Figure 5) with elements as 0s or 1s (black and white image). Another 3x3 matrix (Fig. 6) which is also consisting of 0 and 1s, known as a filter. These two matrices can be considered as two inputs to the convolution function. we get a convoluted matrix that shown in figure 7 After elementwise multiplication by shifting the filter over the original image.

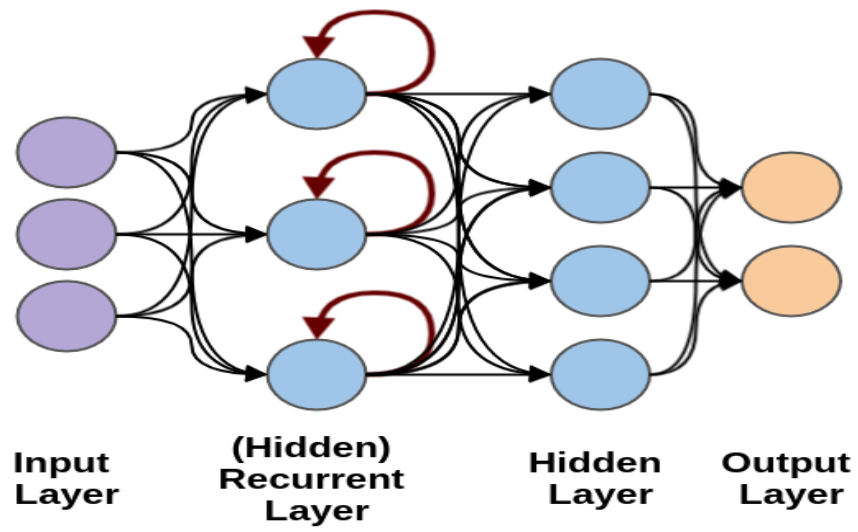


Figure 4: Example of a simple RNN

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

Figure 5: Original image

1	0	1
0	1	0
1	0	1

Figure 6: Convolution filter

- Convolutional Neural Networks:** Convolutional Neural Networks (CNNs) are basically multiple convolution layers which are superimposed with non-linear activation functions such as tanh or ReLU. In general fully connected neural network, every node in the current layer is connected to each node in the next layer. However, in Convolutional Neural Networks convolutional filters are used to navigate through the nodes in the input layer and hence calculate the output. Unlike simple multilayer perceptrons, the concept of sliding convolutions over the nodes in the input layer results in regions of the input layer is connected to every node in the output layer. For computing the output, every layer in the network uses various convolutional filters. When the network processes the training dataset, it automatically learns which filters should be used. See Figure 7 for a simple example of a pooled CNN.

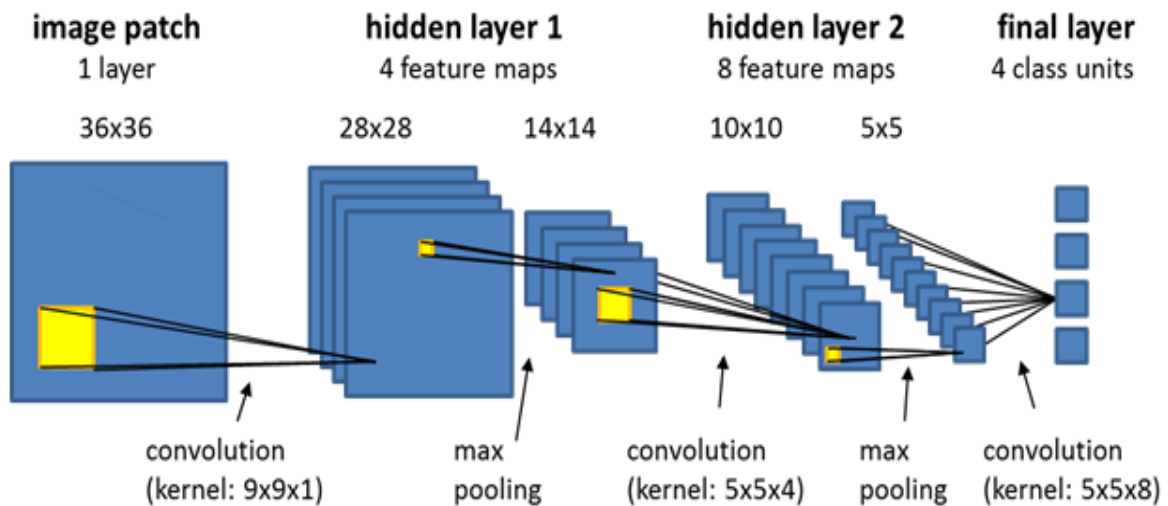


Figure 7: An example of CNN with 1 input layer, 2 hidden layer, 1 output layer, 3 convolution layer, 2 max pooling layer, 4 features map in hidden layer 1, 8 features map in layer 2 and 4 different classes

3.2 Feature Engineering

Feature engineering (FE) is a very important process in the field of data analysis and machine learning in which we need to clean the data and extract the features so that we can obtain better results from our machine learning model. Hence, feature engineering is basically a process of extracting useful features from raw data using mathematics, statistics, and domain knowledge.

3.3 Evaluation Metrics

For any built and trained machine learning model, it is important to evaluate its performance. Different types of models should be evaluated differently. Various metrics are defined to help us assess how well the model exists. The following metrics are most often used for classification tasks: 1) accuracy, 2) precision, 3) recall and 4) score F1.

- **Accuracy:** Accuracy is a metric that typically describes how a model performs across all classes. This is useful when all classes are equally important. It is calculated as the ratio of the number of correct predictions to the total number of predictions.

$$Accuracy = \frac{(\text{No. of true predictions})}{(\text{Total no. of predictions})}$$

- **Precision:** precision is calculated as the ratio of the number of positive samples correctly classified to the total number of samples classified as positive (correct or incorrect). It measures the accuracy of a model in classifying a sample as positive.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall:** Recall is the proportion of correctly predicted positives among all actual positives. The recall measures the model's ability to detect positive samples. The higher the recall, the more positive samples detected.

$$Recall = \frac{TP}{TP + FN}$$

- **F1 Score:** The F1 score is used to measure the performance of the model, considering both accuracy and recall. Mathematically, this is calculated as:

$$F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

4 Existing Techniques and Related Work

4.1 Datasets

Machine learning is a branch of artificial intelligence. A large number of hyperparameter tuning is required to train the model in order to get good results in deep learning algorithms. This hyperparameter tuning happens during the process of training the model when the model is in its learning stage. When data is more to get trained, the model can be more generalized. Thus the good quality of data is very important for implementing any machine learning(ML) and deep learning(DL) algorithm. Most of the dataset was created after 2014. All of these datasets are open for research goals. Some of the datasets are:

- **VQA [9]** : VQA is the most extensively known dataset for solving visual question answering(VQA) problem which was created in the year 2015 as the part of the visual question answering (VQA) competition. They provided the dataset which consists of 265,016 images and 1,105,904 question answer pairs where each image has at least three questions associated with it. The dataset includes balanced real images and abstract scenes. The abstract scenes dataset provides its composition (e.g., left and right facing each clipart object's pixel coordinates) files and captions which can be used for building great models.
- **Visual7W [10]** : This dataset is derived from the MS-COCO (Microsoft Common Objects in Context) dataset which contains 47,300 images and 327,929 question-answer pairs. Each image can have multiple questions associated with it. Visual7W dataset was released in 2016. The question dataset has been created by asking seven "w" questions: why, who, how, where, what, and when. The questions were multiple choice questions and every question had four possible answers. These kinds of datasets were collected by crowdsourcing using Amazon Mechanical Turks.
- **DAQUAR [11]** : The Dataset for Question Answering on Real-world images(DAQUAR) was derived from NYU-Depth V2 dataset in 2015 before the VQA dataset. This is a small dataset that consists of 1500 pictures and auto generated question answer (QA) pairs using nine pre-defined templates.

4.2 Feature Engineering

Feature engineering(FE) is a very important process in the field of data analysis and machine learning in which we need to clean the data and extract the features so that

we can obtain better results from our machine learning model. Hence, feature engineering is basically a process of extracting useful features from raw data using mathematics, statistics, and domain knowledge.

- **Feature Engineering for Images:** To get vectors for images, there are many pre-trained models like VGG-16 [1], ResNet [2], and GoogleNet [3]. Hence, such models are used to extract the features from images and get great results to use in any deep learning model. GoogLeNet[3] was proposed by Google in 2014 which was trained on the ImageNet dataset. In VGG, there are 16-19 layers of convolution neural network and the size of the convolution filter is 3x3.
- **Feature Engineering for Text:** There are many pre-trained models like LSTM, BERT [12], GPT-3 [13], and CodeBERT [?] which are used to extract the features from natural language text data. These models can be further tuned in order to get good performance in natural language processing (NLP) tasks. The pre-trained models are very easy to use and do not require much-labeled data which makes it flexible for solving many business problems.

4.3 Long short-term memory (LSTM):

LSTM is a special version of RNN which solves the short-term memory problem and learns long-term dependencies. Traditional RNNs have short-term memory due to the problem of vanishing gradient so they don't remember what appeared in the beginning of the sentence. LSTMs are commonly used to extract the features from raw text data in visual question-answering problems. The LSTM architecture defines the dimension of word embeddings. LSTM includes three gates called forget gate, input gate, and output gate memory units, called cell states, to solve the short-term memory problem. Finally, I merge the image vector and text vector to get a single embedding.

4.4 Related Work

As we already know, the visual question-answering problem is immensely popular because of promising results of deep learning methods for various tasks related to Computer vision and natural language processing. There are many types of research going on in the direction of proposing the solution for visual question answering. Visual question answering is a problem where different subproblems of natural language processing, computer vision, and knowledge representation need to be performed in order to get succeed. Hence, the proposed model should have a deep understanding of the images and what the question is referring to, and how both image and question

are related.

Kafle et al.[4] proposed a Bayesian model for visual question answering to predict the answer type . the Bayesian model is related to “Quadratic Discriminant Analysis” .To compute feature embeddings for questions, they used skip thought vectors. Skip-thought vectors are very new technique for encoding sentences into vectors in a manner that retains most of the important sentence information.

In 2016, Teni et al.[5] proposed a deep neural network for visual question answering(VQA) that processes graph representations of scenes and question text in which the input image was taken as a complete graph and every node as the object in the image. Edges between every two nodes were taken as the relative position of the objects in the image. This method allows strengthening the existing natural language processing (NLP) like pre-trained word embeddings and syntactic parsing. Words of the question after tokenization were also coded in a graph representing serial relationships and syntactic dependency connections between the words in the question.

Antol et. al.[6], presented a model that uses VGG-16 to extract the features from a given image. Then, the last hidden layer of VGG-16 is l2 normalized and passed it through a fully connected layer to convert into extracted vector. When the entire question has been introduced into the LSTM, it outputs its last state as the question embedding. They utilize a 2-layer LSTM for the question that receives as input the word embedding of each question token, timestep by timestep. Similar to how they project images, this vector (2048 pixels wide) is likewise sent to a fully connected to be projected into the same area. An element-wise multiplication is used to integrate both features for subsequent use by a fully connected layer and a softmax that will forecast the class answer. Here, the classes to forecast are the 1000 most frequent answers in the training set.

5 Proposed Method for Visual Question Answering

We are using the Jupyter Notebook platform for writing code for our visual question-answering project. We are using Microsoft Common Objects in Context (MS COCO) dataset which consists of around 443757 questions, 443757 answers, and 82833 images. This dataset can be downloaded from <https://visualqa.org/download>. Every image can have multiple questions and answers. The question and answer files are available in JSON format in ms coco dataset. Hence, we require to convert the JSON format into a data frame. Here, basic preprocessing is required to do as we do in any NLP task.

We analyzed the dataset to summarise the main characteristic of a given dataset using Exploratory Data Analysis(EDA) in figure 8 and figure 9

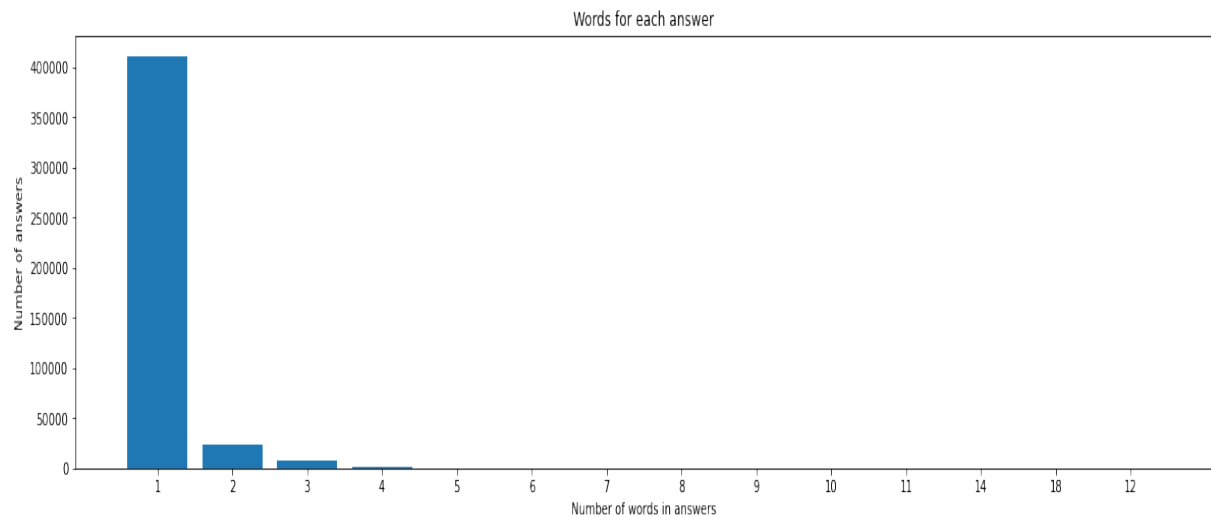


Figure 8: Histogram between No. of words in ans and no. of ans(e.g 1)Is there any humans in the image ? ans:Yes,2)Which sport is being played ? ans: Table tennis

For building the model, we are taking only those questions that have answer types yes or no. We sampled 80000 questions and their answers from 167494 data points. We have done feature extraction from images using a convolution neural network where we first changed the size of an image to (192,192,3). Then , it is passed through different layers of convolution and max-pooling.Hence, we get a vector for an image . Next, it is stored in NumPy format.

Now, using sklearn's train test split, we will split the total data in training and testing

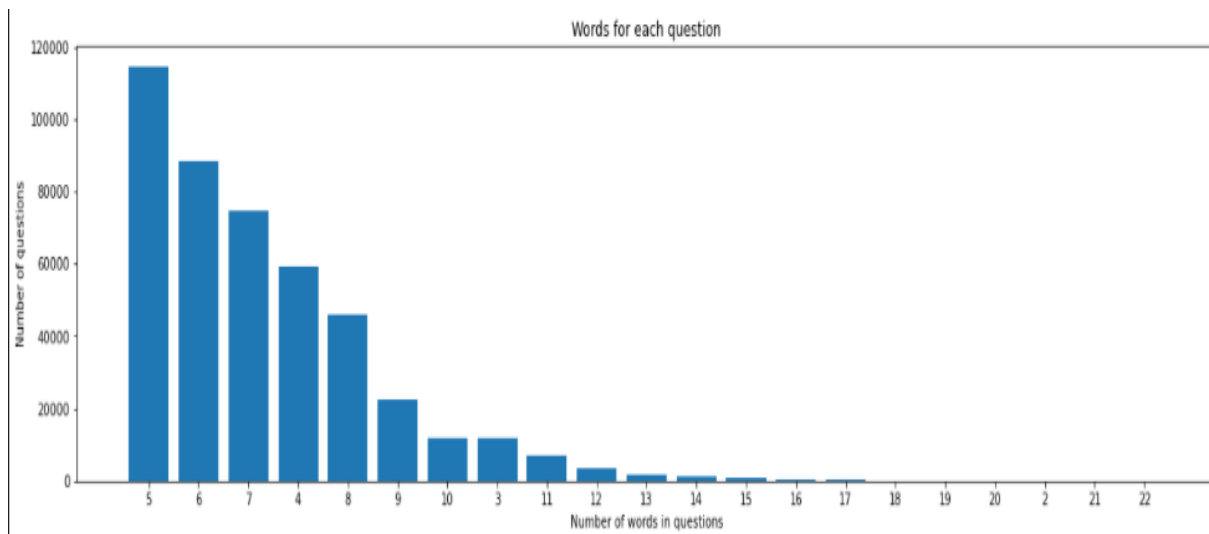


Figure 9: Histogram between no. of words in question and no of question , Minimum and maximum length of question is 5 and 17 words

dataset with the ratio of 70:30. Now we encode the questions and answers. Questions are text functions, so they are encoded using the Keras Tokenizer API [8] where we used GloVe [7] representation. The Keras tokenizer API[8] first breaks the sentence into different words and then assigns each word a unique integer. It is only suitable for train data, so it uses a train data dictionary and does not cause data leakage problems. Each sentence is padded with zeros because each sentence has a different length. Now, each encoded array has the same length. Each word is then represented as a 300-dimensional vector using a pretrained GloVe [7] representation. For answers, we used one-hot encoding where categorical variables are represented as a binary vector.

- **MODEL 1(CNN with Bag of Words(BOW))** The image is sent via CNNs and which gives us a 1024-dimensional vector representing the image. The vector of question text is obtained by averaging the word vectors of all the words present in the question. Then, the image vector and question vector are merged and passed through MLP with two fully connected layers and for regularization, we used dropout(50%). A softmax layer is attached at the end that gives us the output of the answer type(yes or no).
- **MODEL 2(CNN with LSTM)** As we saw that the previous model CNN with BOW ignores the word order in the question. and thus when summing word vectors, The loss of information will be there. For Capturing the sequential nature of the language data, we use LSTM for extracting features from question text. Each word in the question text is first converted to its word embeddings, and these embeddings are transmitted to the LSTM sequentially. That's how we get a 1024

dimension text vector using LSTM for question embedding and then we merged it with an image vector of 1024 dimensions and just apply the same multilayer perceptron architecture which we used in the previous model. See figure 8 for complete architecture

- **MODEL 3(CNN with Bi-LSTM)** For getting more better results, we used bidirectional LSTM for extracting features from the question text and applied the same deep learning architecture that we used in my previous models.

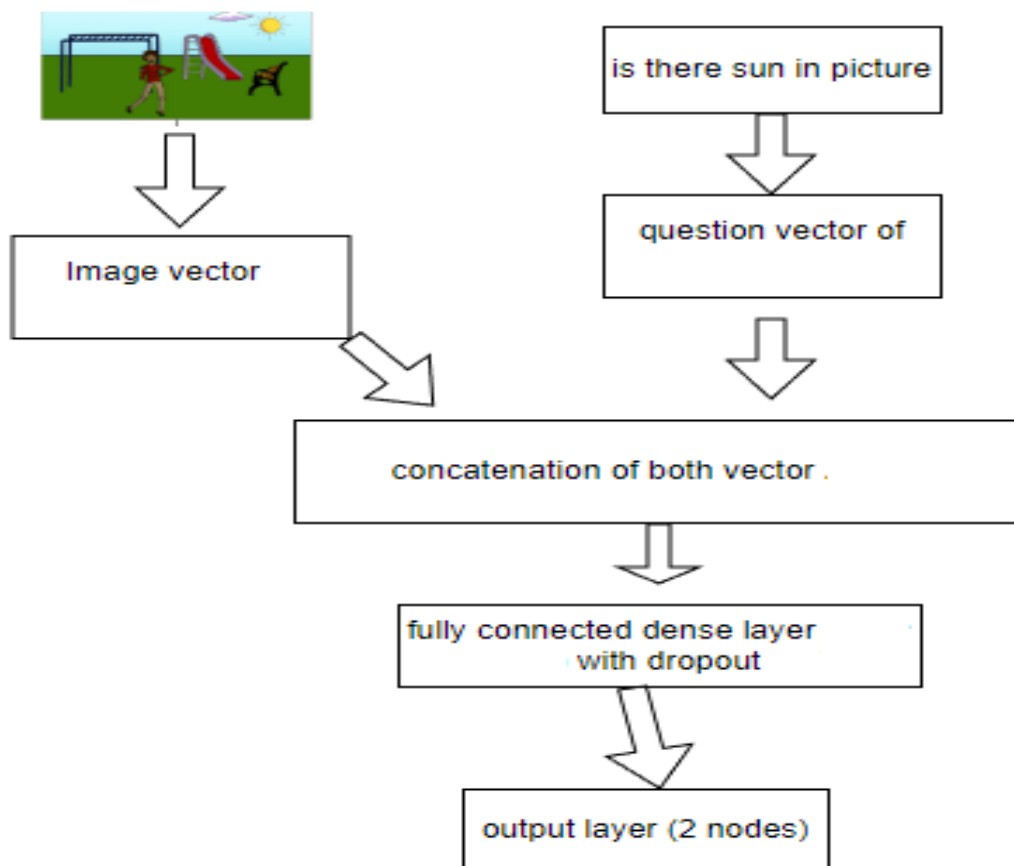


Figure 10: Model Architecture of visual question answering with answer type yes and no

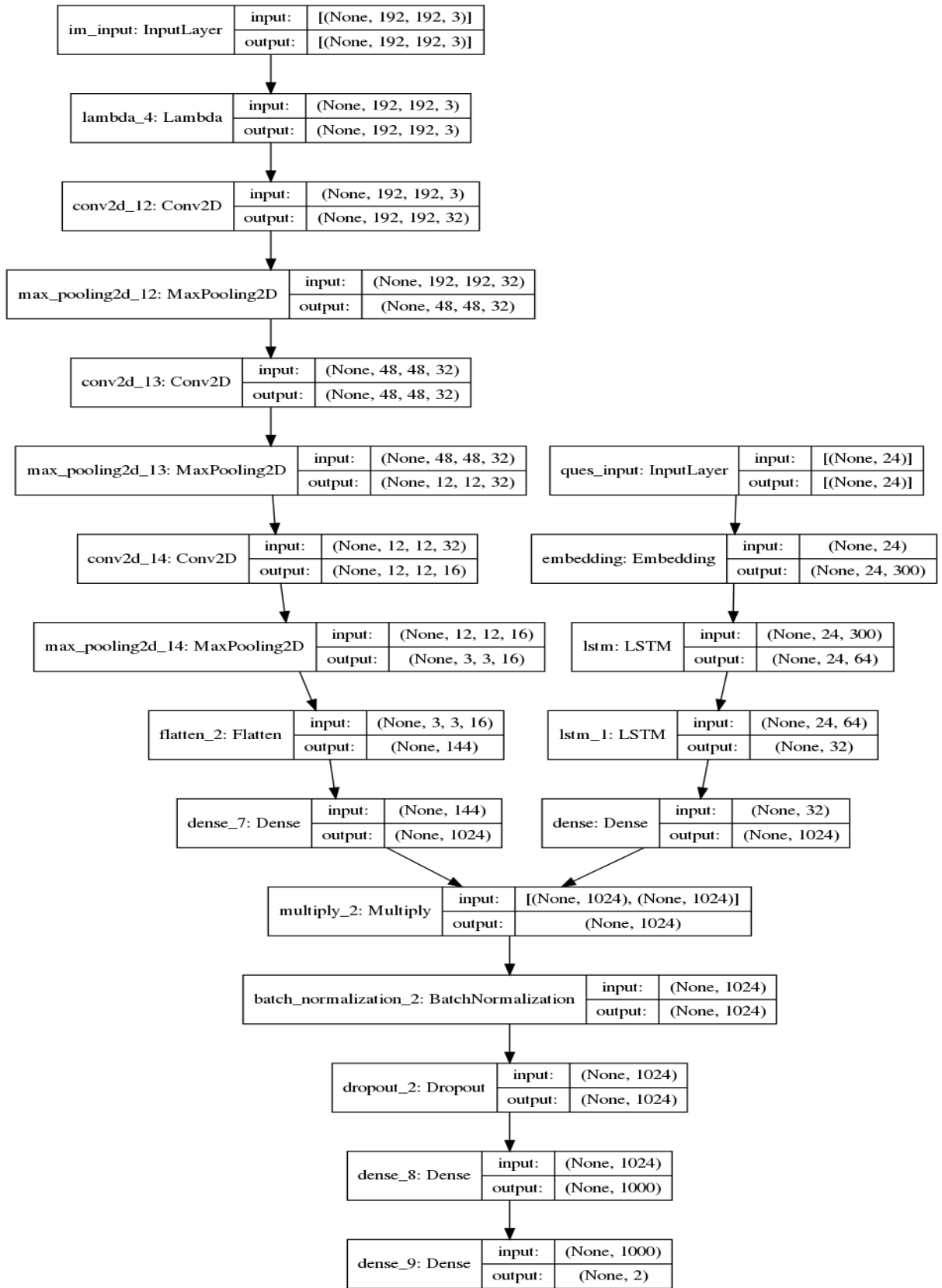


Figure 11: Model architecture for CNN with LSTM

6 Results, Conclusion, and Future Work

- **Results**

Table 1 shows loss and accuracy for all the three models CNN with bag of words, CNN with LSTM, CNN with BI-LSTM on training and testing dataset. We can clearly understand from table 1 that accuracy of model2 improves from model1 because bag of words ignores the word order in the question in model1 and there is loss of information. Whereas LSTM preserves the sequential nature of the question text. Hence model2 performs better than model1.

Model3 improves the accuracy slightly from model2. It happens because unlike standard LSTM, the input flows in both directions, and it's capable of utilizing information from both sides. It's also a powerful tool for modeling the sequential dependencies between words and phrases in both directions of the sequence. Hence model3 performs better than model2.

Model	Model type	Train loss	Train Accuracy	Val loss	Val Accuracy
1	CNN with BOW	0.5334	54.76	0.5545	53.31
2	CNN with LSTM	0.6785	61.17	0.6863	54.74
3	CNN with BI-LSTM	0.6829	59.06	0.6859	55.6

Table 1: Accuracy and Loss

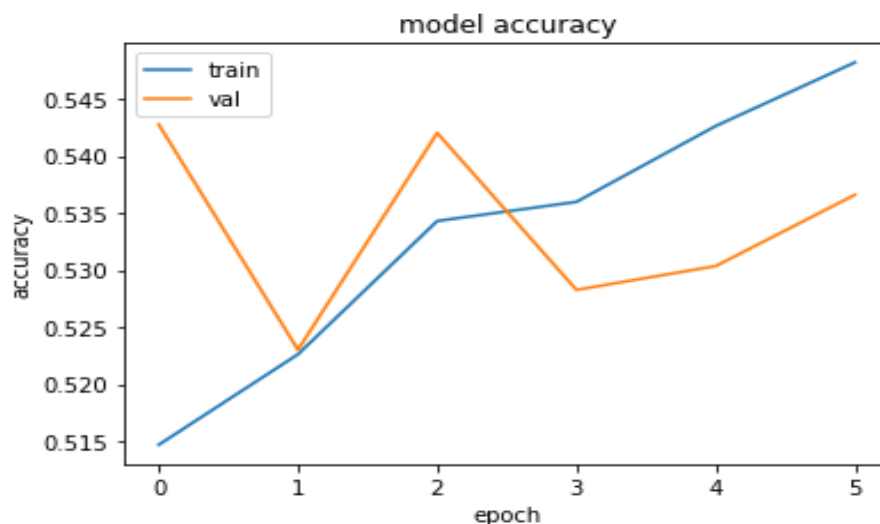


Figure 12: Accuracy vs epoch for model1

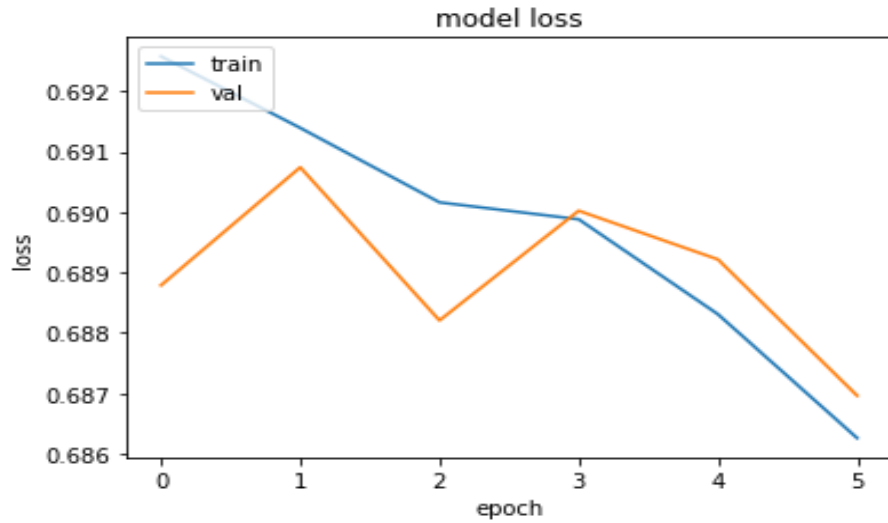


Figure 13: Loss vs epoch for model1

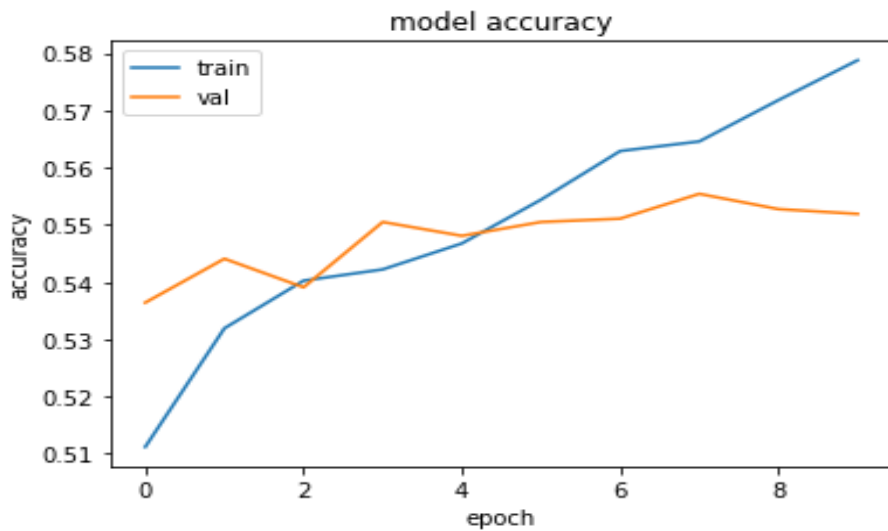


Figure 14: Accuracy vs epoch for model2

- Conclusion** We basically built visual question answering system in this project in which we used mscoco dataset which consists of 443757 questions, 443757 answers, and 82433 images. We framed this as a classification problem where we have to predict the answer from a natural language question about an image. As the dataset is huge, we sample 80000 data points with answers type yes and no. We used convolution neural network for extracting features from image

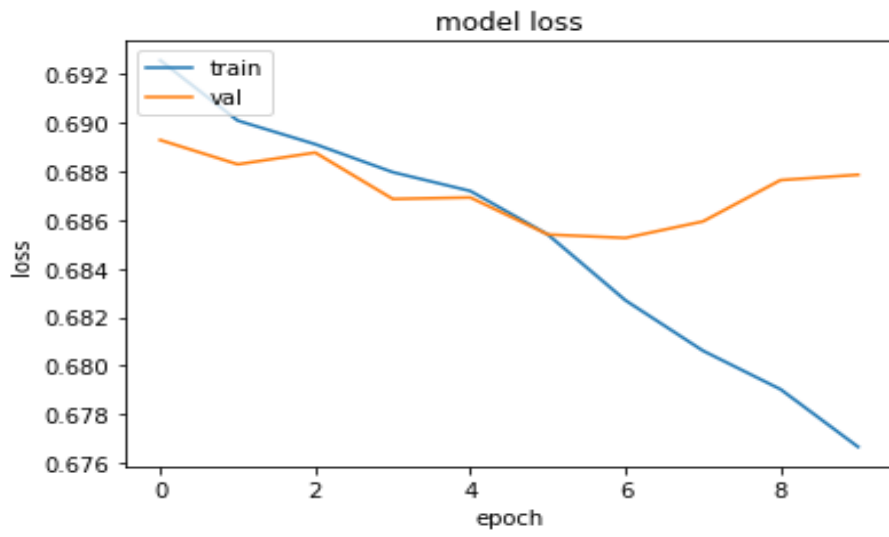


Figure 15: Loss vs epoch for model2

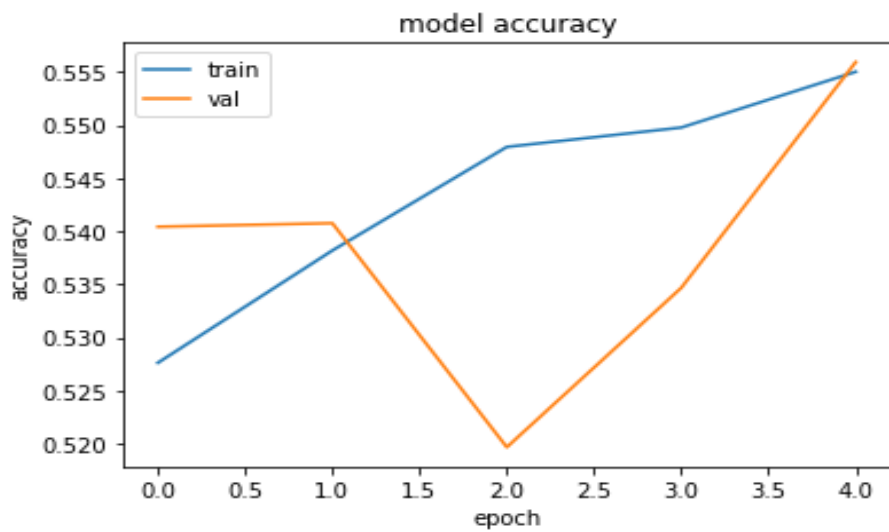


Figure 16: Accuracy vs epoch for model3

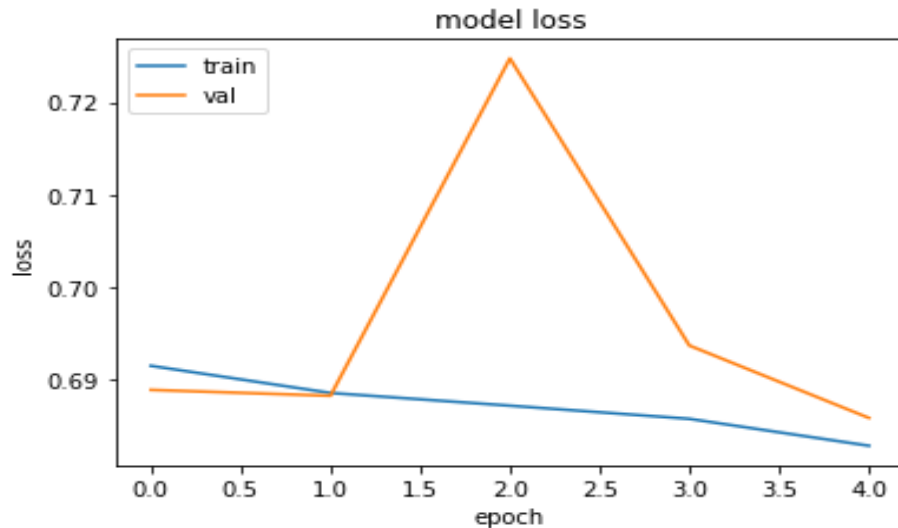


Figure 17: model3

and BOW,LSTM and Bi-LSTM were used for extracting features from question .finally combined them to apply fully connected layer and then softmax layer to retrun an answer as output.



Figure 18: Some examples(1)Is there furniture in the room? Predicted Answer:No,2) Are both people on the bench? Predicted Answer:Yes)

VQA is a relatively new area that requires a deep understanding of both text and image. If we look at the current deep learning and machine learning research scenario, it can be expected that visual question answering systems will definitely improve over time in terms of accuracy and optimality. In recent years,

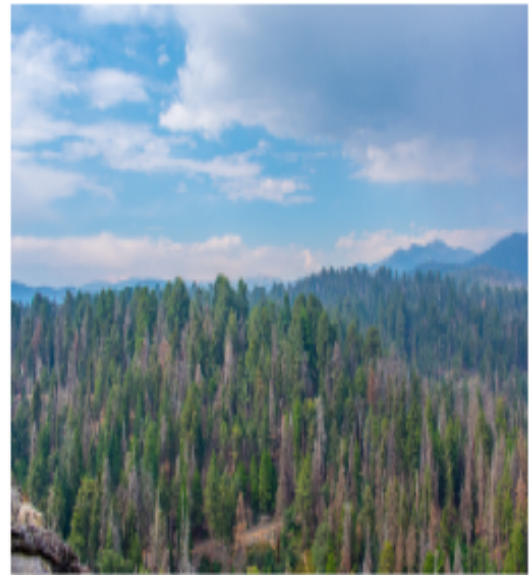


Figure 19: Some examples(1)Is the desk cluttered? Predicted Answer:No,2)Are there clouds in the picture? Predicted Answer:Yes)

there is exponential growth in deep learning performance by various models of computer vision and natural language processing. Using these individual components to build a system that combines them will significantly improve the results are visible for tasks like VQA.

The Accuracy and F1 score are most commonly used metrics in terms of metrics used to measure the performance of a visual question answering. This is due to the fact that all modern systems created for VQA, think of it as a classification problem. They assume that the answer to any a natural language query or question about an image would belong to one of The answer class on which we train the model . One could suggest an extension of this a system that uses a generative model to extract a meaningful response based on on the given image.

- **Future work** The Current modern systems at VQA are still far from human performance on the same data sets. But in the presence of published studies and new technologies and tools, we can expect huge scope for improvement in results.To get better model performance,the large number of hyperparameters in deep learning model are required.High-RAM GPU machine is required for whole big dataset.There are some advanced techniques like stacked attention which can be used in model architechure.

References

- [1] Very Deep Convolutional Networks for Large-Scale Image Recognition Karen Simonyan, Andrew Zisserman
- [2] Deep Residual Learning for Image Recognition Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun
- [3] Going Deeper with Convolutions Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich
- [4] K. Kafle and C. Kanan, "Answer-type prediction for visual question answering," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4976–4984.
- [5] 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 International Conference on Computer Vision, 2015, pp. 2425–2433.
- [6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, pages 2425–2433, 2015.
- [7] GloVe: Global Vectors for Word Representation Jeffrey Pennington, Richard Socher, Christopher D. Manning Computer Science Department, Stanford University, Stanford, CA 94305
- [8] https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/Tokenizer
- [9] <https://visualqa.org/>
- [10] Visual7W: Grounded Question Answering in Images Yuke Zhu, Oliver Groth, Michael Bernstein and Li Fei-Fei
- [11] <https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/vision-and-language/visual-turing-challenge/>
- [12] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
- [13] "Language Models are Few-Shot Learners" Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan