A Reliable Approach for Image Question Answering

Bhagyashree Gawade
University of Central Florida
Orlando, Florida-USA
bhagyashreeg94@knights.ucf.edu

Jidnyesh Sankhe
University Central Florida
Orlando,Florida-USA
jidnyesh.sankhe@knight.ucf.edu

Akshay Dalvi
University Central Florida
Orlando,Florida-USA
akshaydalvi4545@knight.ucf.edu

Manasa Nuthalapati
University Central Florida
Orlando,Florida-USA
manasachowdary@knight.ucf.edu

***Abstract*—Much recent progress in Vision-to-Language problems has been achieved through a combination study of different models comprising the implementation of different Machine Learning algorithms as well as Neural Networks. This work is an attempt to classify and demonstrate how the problem of image-based question answering (QA) which is also refereed to as Visual Question Answering (VQA) can be solved with the help of new models and data sets. In this project, we are using neural networks along with Visual semantic embedding to test the LSTM model with VIS. This model skips the traditional steps of image-question answering like Image segmentation, feature mapping, object detection for predicting the simple questions about particular images. This model outperforms most of the models that are specifically trained for image question answering. It performs 2 times better than the existing models that have worked on a data set that would not be sufficient for training large intricate models. We have used a new data set which is generated using an algorithm that converts image descriptions which are commonly available, into QA form. A collection of results from other baseline models and LSTM+VIS model is also presented by this work.**

I. Introduction

The ability to read a paragraph of text and then answer a question about it i.e. Reading comprehension(RC), is a very challenging task for machines as it primarily requires: The understanding of natural language processing and knowledge of the world. [1] With the study of Natural language processing, the use of computers to answer textual questions started taking pace in the early 1960s. The two major paradigms of question answering were first implemented in 1960. The two major paradigm were: Information retrieval based question answering and Knowledge based question answering. The two early QA systems were BASEBALL and LUNAR. BASEBALL answered questions about the US baseball league over a period of one year. LUNAR, in turn, answered questions about the geological analysis of rocks returned by the Apollo moon missions. Both QA systems were very effective in their chosen domains. In the 1970s, knowledge bases were developed that targeted narrower domains of knowledge. The QA systems developed to interface with these expert systems produced more repeatable and valid responses to questions within an area of knowledge. The current question answering methodologies focus more on the factoid type questions. Factoid questions are questions that can be answered with simple facts expressed in short text answers[2] These questions require an understanding of vision, language and commonsense knowledge to answer. A lot of researchers are now researching on the new form of question answering known as the Visual question answering. The Visual Question Answering is a study which involves research in image and video captioning that combines Computer Vision, Knowlede representation and reasoning and Natural language processing(NLP). In this paper, we have made an attempt to implement the task of open-ended Visual Question Answering (VQA). A VQA system works by first taking an input of an image along with a free-form, open-ended, natural-language question about the image and then goes ahead to produce a natural-language answer as the output. This goal-driven task aims to be utilized in scenarios encountered when visually-impaired users or intelligence analysts actively elicit visual information.[3]

II. Related Work

Recently, there has been tremendous increase in Visual Question-Answering. It is the combination Computer Vision with Natural Language Processing. The progress researchers to build holistic architectures for challenging grounding, natural language generation from image/video image-to-sentence alignment and recently presented question-answering problems. Given an image and the Natural Language question about the image. We must be able to answer the question accurately [4]. It introduces free form and open ended question answering. The VQA system described in the takes the input as the image and the Natural language question and tries to produce the Natural Language answer. Efficient Question Answering requires high A.I. capabilities to answer the fine-grained recognition - what color t-shirt is the man wearing or

Object Detection How many cars are there in the picture. Activity Recognition Is the woman running? Knowledge-Base-Reasoning and Commonsense Reasoning many of the question are mostly yes or no question answering and open ended. As such they have many benefits but it is still required to understand the type of question and the type of answer. One the main requirement of computer vision is scene understanding which further requires the system to understand the objects, actions, events, and scenes. Image Question Answering solves the various recognition problem by unifying the tasks. [5] defines the different recognition task depending on the questions CNN with dynamic parameter layer whose weights are designed adaptively based on the type of question. It helps to build large number weights in dynamic parameter layer effectively and efficiently. Also, GRU is fined trained on large corpus this helps in improving the performance for image Question - Answering. VQA generally depends upon images -which requires to understand of both text (questions) and vision (images). In this case questions are generated by humans making the need for commonsense knowledge and complex reasoning more essential. Also, the task such image captioning, imaging tagging and video captioning. Hence, there is constant focus on where to look in the picture or image for the answer as a human it appears to easy to locate which part of the image to concentrate on is the major concern. [6] It proceeds with image-region selection mechanism which learns to identify the image relevant to the questions. Also, it gives the framework to avoid loss due to the provided baselines. The Model is inspired by the END-to END memory network for question-answering regions in the current model are analogous to sentences in the END-TO-END memory network model many approaches uses recurrent memory network model. Although Simpler Bag of Word (BOW-model) and averaging model roughly outperform the earlier one. we find a fixed-length averaged representation of word2vec vectors for language to be highly effective and much simpler to train, it learns to embed the textual questions and the set of visual image regions into latent space. The input in this model is a question its potential answer and image features from a set of automatically selected regions. Parsed questions and answers are encoded using word2vec and two layered network. Recurrent Neural Networks(RNN) and Long Short Term Memory (LSTM) have proven very effective for general purpose model for concentrating on long term dependences in textual applications. But [7] DMN (Dynamic Memory Network) obtains the high accuracy in the language related tasks. To use DMN as the technique for visual question answering there are few newer modules similar to [5] the input module treats the image region as sentence as text in text module. The input module for VQA is composed of three parts, local region feature extraction- to extract the image feature convolution neural network is used, Visual feature embedding we add the linear tanh activation function to project local regional features, and the input fusion layer- They represent the global features. Question answering model applies to both images and the structural knowledge [8] provides compositional and attentional model to answer the question about images of different type. It proposes two models one collection of neural networks which is freely composed and the second module which standouts as predictor that assembles modules into deep network. This approach has two main contribution first is extension and generalization to attention mechanism, which is the standard tool for manipulating images. Second contribution it assembles the isolated smaller modules into larger structures. It automatically induces variable-free, tree-structured computation descriptors.These are network layouts: they specify a structure for arranging modules (and their lexical parameters) into a complete network.

III. Background

In this report, we have put forth a solution to the problem of a general end-to-end AQ model which uses visual semantic embeddings for building a connection between Convolution Neural Network (CNN) and Recurrent Neural Network (RNN). We also perform comparisons between other baseline models and the LSTM+VIS model on the basis of the accuracy and time each model takes for training and validation. We use an automatic question generating that helps converting description sentences into questions. We do this with the help of a new Dataset (COCO-QA) which was generated in previous theories using an algorithm that follows the same guidelines and a few baseline results on the dataset. Currently, there is a lot of enthusiasm around artificial intelligence, machine learning and deep learning. Deep neural networks have become a dominant modeling tool for various artificial cognition tasks like speech recognition, image recognition, text classification, automatic machine translation, question answer processing, image caption generation, knowledge retrieval, etc.

A. Recurrent Neural Networks

RNNs are most efficient architecture for text processing, language modeling, question answering using. We have seen a lot of architectures for encoding sentences. RNNS are one of them. A recurrent Network can receive a sequence of values as input and can also produce a sequence of values as outputRNN-based models seem to effectively learn representations for information that involves challenges of dealing with variable length text sequences by producing compact representations that will capture long distance relations in the text[25] It is observed that the hidden states are capable of capturing vital data regarding the structure of the input sentences that are efficient and required for performing the predictions. However, it is not an easy task to trace back how this information is captured and what knowledge is obtained from it. As we know RNNS can match quotes and also count parenthesis but it might lack to find out the aspects of languages like grammar, phrases, etc. Hence, our experiments focus on the model that involves LSTM. [25] LSTM defines a variant of the function RNN which has an improved hidden state update to the actual theory. We know that these models are increasingly being used for different purposes. LSTMs and RNNs can be stacked in layers for developing multiple hidden

state vectors at each time step, which indirectly enhances the performance.

B. Convolution Neural Networks

The vision part extracts visual features through a deep convolutional neural network (CNN). [26] or other way is using a traditional visual feature extractor. The question understanding part includes a dense question embedding feature vector to encode question semantics, either with a Bag-of-Words model or a recurrent neural network (RNN) model. Lastly, the answer generation part generates an answer conditioned on the visual features and the question embeddings.[27]

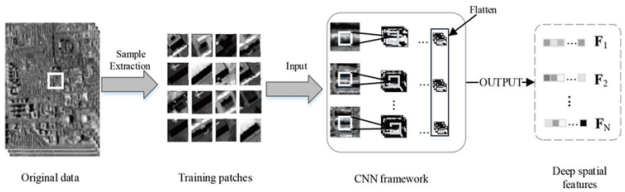

Fig. 1. Feature extraction with a CNN framework. Training samples are extracted from the original data. A CNN framework is trained with training samples. The features at the last layer of the CNN framework are flattened to form the feature vectors.

1) Influence of Image CNN and Competence of Sentence CNN:: The accuracy of human answering the questions falls from 50% to 12% without the image content[28] Hence, it is very important for the image question answering task. As mentioned in the work (Malinowski, Rohrbach, and Fritz 2015; Ren, Kiros, and Zemel 2015), we have also used the question representation that is obtained from the sentence CNN to predict the answer. Most approaches based on deep learning commonly use CNNs to extract features from image while they use various other guidelines to handle question sentences. Some algorithms employ embedding of joint features based on image and question [29]. However, learning a softmax classifier on the simple joint featuresconcatenation of CNN-based image features.

IV. Dataset

At the beginning of 2014, the datasets used for performing Visual Question Answering were : DAQUAR [9], COCO-QA [10], The VQA Dataset [11], FM-IQA [12], Visual7W [13], and Vi- sual Genome[9]. All of the above major image question answering datasets, except the DAQUAR dataset are made up of the the Microsoft Common Objects in Context (COCO) dataset. The Microsoft Common objects in context dataset (COCO) consists of 328,000 images, 91 common object categories with over 2 million labeled instances, and an average of 5 captions per image. An ideal VQA dataset needs to be sufficiently large to capture the variability within questions, images,and concepts that occur in real world scenarios. It should also have a fair evaluation scheme that is difficult to game and doing well on it indicates that an algorithm can answer a large variety of question types about images that have definitive answers. If a dataset contains easily exploitable biases in the distribution of the questions or answers, it may be possible for an algorithm to perform well on the dataset without really solving the VQA problem. [14] The two main datasets that are usually utilized by researchers for Visual question answering are the DAQUAR dataaset and the COCO-QA dataset. We will review the two dataset to understand them better.

A. COCO-QA dataset

The common objects in context (COCO) is a new image recognition, segmentation, and captioning dataset. COCO has several features: Object segmentation Recognition in Context Multiple objects per image More than 300,000 images More than 2 Million instances 80 object categories 5 captions per image Keypoints on 100,000 people The COCO-QA dataset consists of Question Answering pairs that are created for images with the help of an algorithm based on Natural Language Processing (NLP) that derive the QA pairs from the image captions. Fo example, if a image is captioned with the sentence, The milk is stored in the fridge, a question can be created based on the caption such as ,where is the milk kept? along with its answer that it is stored in the fridge. The dataset consists of major part of the question focussing on the object in the image(69.84%), while the other focus on questions about color (16.59%), counting (7.47%) and location (6.10%). All of the questions have a single word answer, and there are only 435 unique answers. These constraints on the answers makes evaluation relatively straightforward. COCO-QA contains 78,736 training and 38,948 testing QA pairs. Although the COCO-QA dataset is widely preferred for Visual Question answering, it has some shortcoming. The shortcomings are due to the flaws in the algorithm based on Natural Languae Processing(NLP) that was used to obtain the Question Answers pairs. Longer sentences are broken into smaller chunks for ease of processing, but in many of these cases the algorithm does not cope well with the presence of clauses and grammatical variations in sen- tence formation. This results in awkwardly phrased questions, with many containing grammatical errors,and others being completely unintelligible. The other major shortcoming is that it only has four kinds of questions, and these are limited to the kinds of things described in COCOs captions. [14]

B. DAQUAR

The DAQUAR dataset stands for The DAtaset for QUestion Answering on Real-world images and it was the first and a comparatively smaller dataset. The images from the NYU-Design v2 dataset helped to build the dataset. The DAQUAR dataset contains abot 1449 RGBD images of indoor scenes, which also includes the annotated semantic segmentations. Two types of question answer pairs are collected. First, synthetic questions answers are generated auto- matically using 8 predefined templates and the existing annotations of the NYU dataset. Second,human questions answers are collected from 5 annotators. They were instructed to focus on basic colors, numbers, objects (894 categories), and sets of those. Overall,

12,468 question answer pairs were collected, of which 6,794 are to be used for training and 5,674 for testing. The images of DAQUAR are split to 795 training and 654 test images. The main disadvantage of DAQUAR is the restriction of answers to a predefined set of 16 colors and 894 object categories. The dataset also presents strong biases showing that humans tend to focus on a few prominent objects, such as tables and chairs.

V. Approaches

Answering open-ended questions is a fundamental ability for any intelligent framework. A standout amongst is recent open-ended question answering challenge is Visual Question Answering which attempts to evaluate a systems visual comprehension by delivering answers to natural language questions about images. There exist many approaches to VQA, the majority of which do not exhibit deeper semantic understanding of the candidate answers they produce. First attempt at open-world VQA included combining of semantic text parsing with image segmentation in Bayesian formulation that samples from nearest neighbors in the training set which required human-defined predicates which were dataset specific and difficult to scale along with dependency on accuracy of image segmentation algorithm and of the estimated image depth information. Another attempt was based on a joint parse graph from text to videos and hence made approaches restrict questions to predefined forms. Later, different methods were introduced which are categorized as follows: joint embedding approaches, attention mechanisms, compositional models and knowledge base-enhanced approaches.

A. Joint Embedding approaches

A method called Neural-Image-QA was proposed with a Recurrent Neural Network (RNN) implemented along Long Short-Term Memory cells (LSTMs). The purpose behind RNNs is to handle questions and answers i.e. inputs and outputs of variable size. Several variants of this approach were proposed. For example, VIS+LSTM-main focus of this paper, formulated the answering as a classification problem whereas the former variant treated it as a sequence generation procedure. Another variant with slight technical enhancement was proposed: 2-VIS+BLSTM model which used two sources of image features as input (fed to the LSTM at the start and end of the question sentence) and LSTMs that scanned questions in both forward and backward directions which helped in capturing better relations between distant words in the question.[33] Many other methods such as Multimodal QA (mQA), CNN with dynamic parameter layer (DPPnet), Multimodal Compact Bilinear pooling (MCB) , multimodal residual learning framework(MRN) , DualNet were proposed with combined multiple strategies.

1) Performance: MCB and MRN methods achieve top performances and show scope for further research and improvement. These approaches constitute the base for most current approaches to VQA.

B. Attention Mechanisms

The main aim of attention mechanism is to address issues by using local image features and allowing the model to assign different importance to features from different regions. The attentional component of the model identifies salient regions in an image and further then focuses the caption generation on those regions. A method that described how to add special attention to the standard LSTM model was proposed.[33] Another compositional model that builds a neural network from modules tailored to each question came into emergence. Most of these modules operate in the space attentions, either producing an attention map from an image, performing unary operations or iterations between attentions.

1) Performance: Attention mechanisms have reported improvement in models that use global image features. Attention-enhanced LSTM outperforms VIS+LSTM model and this mechanism improves the overall accuracy on all VQA but limited to question types.

C. Compositional Models

This approach involves connecting distinct modules designed for particular desired capabilities such as memory or specific types of reasoning with a better use of supervision. It facilitates transfer learning as a same module can be used and trained within different overall architectures and tasks. On the other hand, it allows to use deep supervision i.e. optimizes an objective that depends on the outputs of internal modules. This model mainly focuses on two models Neural Module Networks (NMN) and Dynamic Memory Networks (DMN).[33] One of the markable contributions of NMNs is to apply logical reasoning over continuous visual features instead of discrete or logical predicates. DMNs is designed to address tasks that require complex logical reasoning by modeling interaction between multiple parts of the data over several parses.

1) Performance: The DMNs were evaluated on the DAQUAR and VQA datasets to show competitive results for all types of questions. Compared to NMNs, the performance was on similar grounds for yes/no questions, slightly worse on numerical questions but reasonably better on all other types of questions.

D. Models using external knowledge bases

A substantial amount of research on structured representations of knowledge led to development of Knowledge Bases (KB) such as DBpedia, Freebase etc., stores common sense and factual knowledge in a machine readable form. A VQA framework called Ahab was proposed that uses DBpedia.[33] Later, an improved method called FVQA which uses LSTM and data-driven approach to learn the mapping of images/questions to queries emerged. A joint embedding approach that benefits from external knowledge bases was proposed.

1) Performance: In terms of overall accuracy, Ahab outperforms joint embedding approaches on KB-VQA dataset. FVQA shows better results than conventional approaches in terms of overall top-1 accuracy. A joint embedding approach benefiting from external KS shows advantage in terms of average accuracy on COCO-QA and VQA datasets[33]

VI. Proposed Methodology

The development of the model is based on the applications of various forms of neural networks and visual-semantic embeddings. Recurrent Neural Networks and to be specific, long short-term memory networks (LSTMs), are an efficient tool for modelling sequentially that implements a dense black-box hidden representation of their sequential input. Recently, a lot many theories of layered, end-to-end trainable artificial neural networks have made improvements in the performance over an extensive range of varied tasks. [15] In case of larger datasets, deep convolutional neural networks (CNNs) compete the capabilities of humans to conduct image classification. [K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in CVPR, 2016.]. We can clearly state that Long Short Term Memory Networks are dominating results on a range of sequence prediction tasks like machine translation. These two theories of employing neural architectures have been combined profitably with methods which generate image [16]and video specifications. Both these ideas are conditioning on the visual structures that stem from deep learning architectures and employ recurrent neural network approaches to produce descriptions. [17] In this work, we have tried representing how LSTM+VIS as an analysis technology for recurrent neural networks with the idea of comprehending the hidden state dynamics. This technology helps the user to choose hypothetical input range to concentrate on the state changes. So that, these changes will match the pattern that are similar in the larger datasets and also align the results and analysis with domain specific mechanical annotations. The following image describes the approach of Neural-Image-QA to question answering with a Recurrent Neural Network using Long Short Term Memory (LSTM). The image that is CNN features and the related question are both fed into the LSTM. Then t rounds of history are considered which are also given to LSTM. The inner product then goes to network. After encoding the questions, the answers are generated. In the answer generation phase, the previously predicted answers are also fed into the LSTM. Then the decoder predicts the actual answer with the reference answers with an accuracy.

LSTM unit: The figure 2. Explains the LSTM unit that takes input vector vt at each time step t and predicts an output word zt that will be equal to its latent hidden state ht. zt is a linear embedding of the corresponding answer word at. The LSTM unit additionally maintains a memory cell c. This helps to learn long-term dynamics more easily and meaningfully reduces the vanishing and exploding gradients problem [18]. We use the LSTM unit as described in and the Caffe implementation from [5]. With the sigmoid nonlinearity : R 7 [0, 1], (v) = (1 + e v) 1 and the hyperbolic tangent nonlinearity : R 7 [1, 1], (v) = e ve v e v+ev = 2(2v) 1, the LSTM updates for time step t given inputs vt, ht1, and the memory cell ct1 where denotes element-wise multiplication. All the weights W and biases b of the network are learnt jointly with the cross-entropy loss. [19]

A. Model 1: VIS+LSTM and Model 2: 2-VIS+LSTM

This project has 2 models VIS+LSTM and 2-VIS+LSTM which are versions of VSE model. VIS+LSTM has a single LSTM for encoding the image and question in single direction while 2-VIS+LSTM uses a bidirectional LSTM to encode the image and questions along with both the directions to take help of the interactions between the image and each word of the question. It is observed that 2-VIS+LSTM performs better than 2-VIS+LSTM with a bigger margin. We have evaluated and compared both the models on the COCO-QA dataset. The table illustrates that the bidirectional LSTM can efficiently model the interactions between image and questions than a single LSTM. Recently, Recurrent Neural Networks have succeeded in the areas of natural language processing. The model VIS+LSTM treats image as a single word of the question. We compare VIS+LSTM with other baseline models like Bag-of-words (BOW), IMG+BOW, LSTM in the Experimental Results section.

1. In this paper, as our visual embeddings we make use of utmost hidden layer of the 19-layer Oxford VGG Conv Net trained on ImageNet 2014 Challenge. During the training time, Convolution Neural Network is kept constant. [20] 2. The dataset has been experimented with various word embedding models: randomly initialized embedding, general-purpose skip-gram embedding and dataset-specific skip-gram embedding model. Along with the rest of the model, word embeddings are trained. [21] 3. The model then considers image as first word of the sentence. To map 4096-dimension image feature vectors to a 300/500 dimensional vector a linear or affine transformation is used that matches the dimension of the word embeddings. [22] 4. The image can optionally be treated as end word of the question as well through a different weight matrix and electively add a reverse LSTM which gets the matching content but runs in a backward sequential manner. 5. At the last timestep, LSTM/LSTMs outputs are fed into a soft-max layer to generate answers.

Our proposed CNN model significantly outperforms the competitor models. More specifically, for the case of single word, our proposed CNN achieves nearly 20% improvement in terms of accuracy over the best competitor model 2-VIS+BLSTM

VII. Algorithm used for VQA

Abundant VQA algorithms have been studied till now. In general, a VQA algorithm consists of three main components: 1) Image featurization, 2) Question featurization, and 3) A methodology to process these features to produce an answer. There are various alternative formulations (e.g., [23], the common VQA systems describe this problem as a classification problem in which the system is provided an image and a

question, with the answers are provided as categories. Image and question featurization are common amongst technologies irrespective of the type of arrangement utilized for the generation of answers. Most of the algorithms use CNNs that are pre-trained on ImageNet. There are a variety of techniques used for extraction of text features such as bag-of-words (BOW), long short term memory (LSTM) encoders [24]. For answer generation, the generic methodology is to treat VQA as a classification query.

VIII. Importance of Visual Question Answering

1) Making the V in VQA Matter: Enriching the Role of Image Understanding in Visual Question Answering 2)

Fig. 2.

Balancing and Answering Binary Visual Questions 3)The

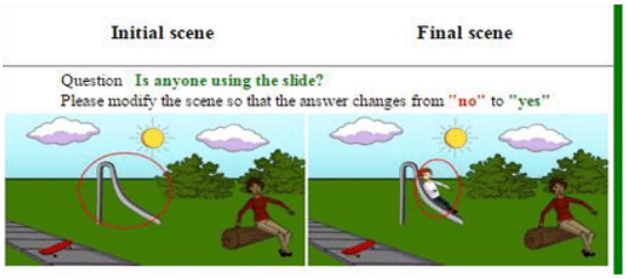

Fig. 3.

above figure shows how image with a question allows the model to make quicker predictions.

Dataset	Input	All	Yes/No	Number	Other
Real	Question	40.81	67.60	25.77	21.22
	Question + Caption*	57.47	78.97	39.68	44.41
	Question + Image	83.30	95.77	83.39	72.67
Abstract	Question	43.27	66.65	28.52	23.66
	Question + Caption*	54.34	74.70	41.19	40.18
	Question + Image	87.49	95.96	95.04	75.33

Fig. 4.

IX. Question Answer Generation

One of the important tasks of image question answering in Natural Language processing is learning to ask questions. Being certain on what to question about indicates understanding and as same, generating questions marks machine understanding. There are many datasets available currently such as DAQUAR which contains around 1500 images and 7000 questions on 37 common object classes. [34] This dataset does not yield reasonable results on training large complex model and random guessing leads to higher accuracy. This problem was overcome by using another data set Microsoft Common Objects in Context (MS-COCO) which includes day scenes with 91 basic objects in 328k images, each paired with 5 captions. [35] This data set automatically converts image descriptions such as image labeling into QA form which helps in relying more on image understanding. Such a conversion also maintains language variance in original labeling giving more human-like questions than image description generated questions. In this project, we use COCO-QA which generates questions automatically from image descriptions of MS COCO dataset by applying a set of transformation rules to generate wh-question. [35]

A. Question Generation

Question generation for QA purpose is still an open-ended topic. Type of question posed plays a significant role. There are huge range of topics for question generation but narrowing down to this project, we consider generating four types of questions as given: Object Questions: At first, we use what for asking about an object. This includes replacing the actual object with a what and then changing the sentence structure such that what appears at begin of the phrase/sentence.

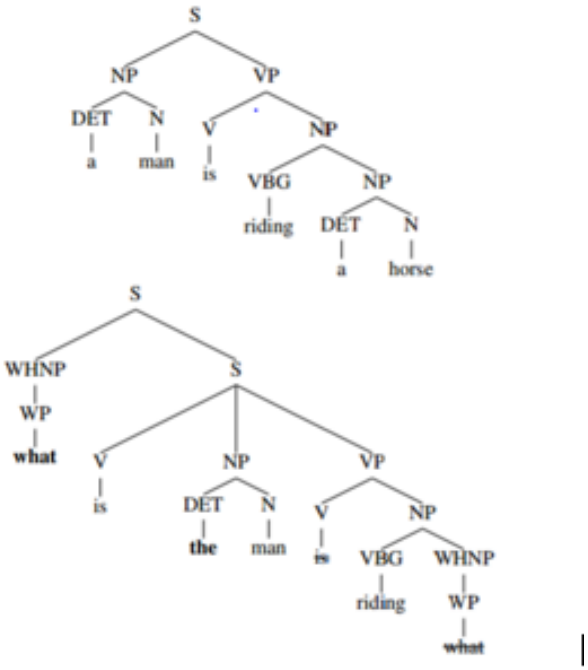

Fig. 5.

Number questions: These include How many type of questions which basically gives the count from the original sentence. Color questions: These type of questions are quite easy to generate. This requires just locating the color adjective and the noun which the adjective attaches to. Next, it forms a sentence What is the color of the object replacing the actual noun. Location Questions: A similar type of question generation as object questions except for that now the answer traversal will only search within preposition (PP) constituents that begin with the preposition in. Filtering was done to main stream only locations, scenes or large objects which contain smaller ones.

X. Common Strategies for VQA

We utilize few strategies which proved to be useful in forming questions 1. Compound sentences to simple sentences In this strategy, we only consider a straight forward case where two sentences are combined together with a conjunctive words. The original sentence is divided into two independent sentences. For example: There is a dog and the dog is running will be split to There is a dog and the dog is running 2. Indefinite determines to definite determiners Changing the determiner into definite form the is a must for posing questions on particular instance of the subject. For example: A girl is watching TV. Will have the instead of a in its question form: What is the boy playing? 3. Wh-movement constraints In general, questions tend to start with interrogative words such as what, who etc. The algorithm which was used for dataset needs to move the verb as well as the wh-constituent to the front of the sentence. For the project purpose we take into account two simple constraints: (a) A-over-A principle The A-over-A principle restricts wh-word movement inside a noun phrase (NP). For example: I am talking to Sam and Phill is not transformable to Who am I talking to Sam and as Phill is a NP which is under a different NP Sam and Phill (b) Clauses Movement of wh-word containing in the clause constituent is restricted.

XI. Post-processing

The COCO-QA dataset used eliminates the answers that appear rarely or not too often. Answers that show less than a threshold are discarded. For the dataset used, threshold is around 20 in training set and 10 in test set. Next, all the QA pairs are randomized to eliminate the dependencies between neighboring pairs. We develop the rejection process as Bernoulli random process. The probability of accepting the next QA pair (q,a) is:

$$p(q,a) = \begin{cases} 1 & \text{if count}(a) \leq K \\ \exp\left(-\frac{count(a)-K}{K_2}\right) & \text{otherwise} \end{cases}$$

Fig. 6.

Where count(a) denoted the current number accpeted QA pairs that have a as the ground truth answer and K,K2 are some constants with K¡K2. In the COCO-QA generation we chose K=1000 and K2 = 200. After the elimination, the dataset has uniform distribution across the possible answers.

XII. Experimental Setup and Implementation

This project is basically implanted to learn image question answering in detail along with the different models and datasets. We have also made a comparative study of various models implemented previously for VQA. This project is implemented in the Keras which is a high level Neural Network Application Programming Interface. We have used Python programming language. We have implemented and tested VIS+LSTM image question answering model. This model is explained in the report. Another model is tested which is 2-VIS+LSTM model. This model has 2 image feature inputs, at the start and at the end of each sentence along with different learned linear transformations. We have worked on a pre-processed dataset which is COCO-QA. The processing of this dataset is available on the internet.

A. Requirements of the project

Python 2.7 Numpy Scipy (to load pre-computed MS COCO features) NLTK (to tokenize) Keras Theano

B. Training of the Models

The 2 models are trained and validated on the COCO-QA dataset. We run the train.py file by specifying the batch size and the number of epocs using the options num_epocs and batch_size. The default batch size can be 200 and the number of epocs can be less than or equal to 25. We have trained the model changing the epocs from 10 to 25. For instance, to train 2-VIS-LSTM with a batch size of 50 and for epocs 10, we can use: python train.py model=2 batch_size=50 num_epocs=10. We have trained the model on different Computers. So, there could be a memory issue because of the RAM and hence the batch size has to be reduced from 200 to 100 or 50.

C. Models

The working of the models is described in the following figures. 1. VIS+LSTM:

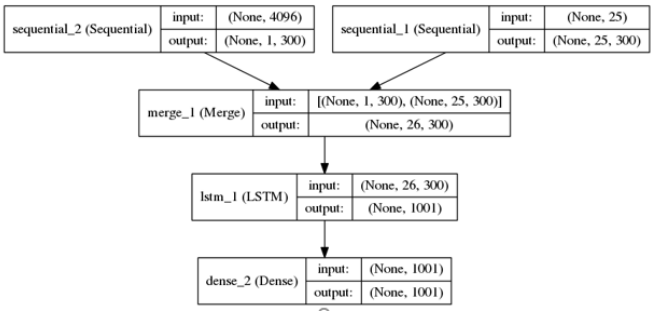

Fig. 7.

2. 2-VIS-LSTM:

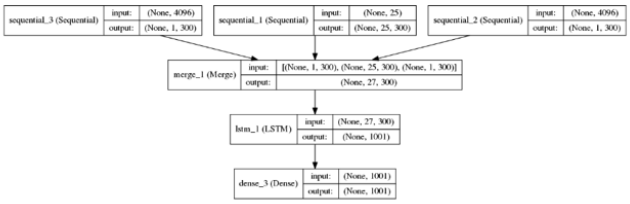

Fig. 8.

D. Predictions

Question answering can be performed on the image using the following script: Question_answering.py The options question and image are used to specify the question and the path the image. The model that you want to use for prediction is specified using the script model. By default model 2 is selected. For instance, python question_answer.py image=example.jpg question=Who is reading the book? model=2

XIII. Results and Analysis

Fig. 9.

We asked the question, How many slices are there in the pizza?

The answer returned was 6.

Refer Fig 10. The question asked was, What is the sport being played?

The answer reurned was Cricket.

After successfully training the two models using python implementation. We then evaluate and compare its results along with the existing datset. Based on the batch size and the number of epochs, the training model gave different results for the loss and validation accuracy.

The fig 12 identifies that the accuracy for the models is quite impressive as we also have a low loss function. We also realize that increasing the batch size and the Number of epoch increases the validation accuracy. Although, Increasing the

Fig. 10.

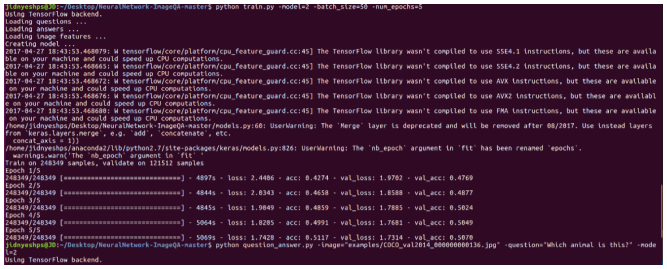

Fig. 11.

number of epochs or the batch size increases the computational time. We then performed a comparative study basd on the available results to compare the accuracy of our model as compared to the existing implemented model gave us a good insight about the usefulness of our model in the current scenario of Visual question answering.

From the fig 13 we see that the BOW model gives an accuracy of 37.52% for the COC-QA dataset whereas it gives 32.67% for the DAQUAR data. The original LSTM implementation with the COCO-QA and DAQUAR dataset gives an accuracy of 36.76 and 32.73% respectively. Using an extension of the BOW by combining the IMG algorithm gives a better result of 55.92% with the COCO-QA dataset. From the above results we observe that our model outperforms the baelines and the existing approach in terms of answer accuracy. We also compared te accuracies after using the DAQUAR dataset to confirm that the choice of COCO-QA dataset for he Visual question answering purpose was the right one. Using the two types of VIS+LSTM models with different word embeddings techniques to check for accuracy did not differ much in accuracy. Both the model showed similar accuracies with a similar loss function as well.

Model	Batch size	Epoch	Accuracy	Loss
VIS+LSTM	200	10	53.27%	1.34
VIS+LSTM	50	5	50.34%	1.41
2VIS+LSTM	200	10	54.01%	1.45
2VIS+LSTM	50	5	51.17%	1.52

Fig. 12.

Model	COCO-QA accuracy	DAQUAR accuracy
BOW	37.52%	32.67%
LSTM	36.76%	32.73%
IMG+BOW	55.92%	34.17%
VIS+LSTM	53.27%	34.12%
2VIS+LSTM	54.01%	35.44%
K-NN	44.96%	31.85%

Fig. 13.

XIV. Conclusion

We implemented the Visual Question Answering problem using end-to-end neural network models. Although, the topic is nave and still under development we have achieved reasonable understanding of the questions along with some image understanding. Neural networks are recently gaining popularity and are considered most suitable for Image and text processing. Bag-of-word model(BOW) performs equally well and can be compared with recurrent networks that is borrowed from image captioning generation. Image question answering is considerably new topic in area of research leveraging techniques from computer vision, artificial intelligence and natural language processing. Most question we analyzed are YES or No questions. We would like to consider the model for longer answers. In this paper, we focus on Limited domain questions. Visual attention is another future work which can aid to improve the results as well as help in model prediction by examining the attention as output at every time-step.

References

[1] [1] Pranav Rajpurkar and Jian Zhang and Konstantin Lopyrev and Percy Liang,SQuAD: 100,000+ Questions for Machine Comprehension of Text

[2] [2] Daniel Jurafsky James H. Martin, Speech and language processing, Question answering.

[3] [3] Daniel Jurafsky James H. Martin

[4] [4] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh, VQA: Visual Question Answering.

[5] [5] Hyeonwoo Noh Paul, Hongsuck Seo, Bohyung Han, Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction.

[6] [6] Kevin J. Shih, Saurabh Singh, and Derek Hoiem, Where To Look: Focus Regions for Visual Question Answering.

[7] [7] Caiming Xiong*, Stephen Merity*, Richard Socher, Dynamic Memory Networks for Visual and Textual Question Answering.

[8] [8] Jacob Andreas and Marcus Rohrbach and Trevor Darrell and Dan Klein,Learning to Compose Neural Networks for Question Answering.

[9] [9] M. Malinowski and M. Fritz, A multi-world approach to question answering about real-world scenes based on uncertain input, in NIPS, 2014.

[10] [10] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, VQA: Visual question answering, in ICCV, 2015.

[11] [11] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang,and W. Xu, Are you talking to a machine? Dataset and methods for multilingual image question answering, in NIPS 2015.

[12] [12] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei,Visual7w: Grounded question answering in images, in CVPR, 2016.

[13] [13] R. Krishna, Y. Zhu, O. Groth, J. Johnson,K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei, Visual Genome: Connecting language and vision using crowdsourced dense image annotations, 2016.

[14] [14] Kushal Kafle1 and Christopher Kanan1*1Chester F.Visual Question Answering: Datasets, Algorithms, and Future Challenges

[15] [15] Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks: Hendrik Strobelt, Sebastian Gehrmann, Bernd Huber, Hanspeter Pfister, and Alexander M. Rush Harvard School of Engineering and Applied Sciences1 hstrobelt, gehrmann, huber, pfister, rush@seas.harvard.edu, 2015

[16] [16] Ask Your Neurons: A Neural-based Approach to Answering Questions about Images, Mateusz Malinowski1 Marcus Rohrbach Mario Fritz Max Planck Institute for Informatics, Saarbrucken, Germany 2UC Berkeley EECS and ICSI, Berkeley, CA, United States, October 1, 20154.

[17] [17] I. Sutskever, O. Vinyals, and Q. V. V. Le. Sequence to sequence learning with neural networks. In NIPS. 2014.

[18] [18] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In CVPR, 2015

[19] [19] Ask Your Neurons: A Neural-based Approach to Answering Questions about Images, Mateusz Malinowski1 Marcus Rohrbach Mario Fritz Max Planck Institute for Informatics, Saarbrucken, Germany 2UC Berkeley EECS and ICSI, Berkeley, CA, United States, October 1, 2015

[20] [20] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Computation, 1997. [10] M. Iyyer, J. Boyd-Graber, L. C

[21] [21] W. Zaremba and I. Sutskever. Learning to execute. arXiv preprint arXiv:1410.4615, 2014.] . [Ask Your Neurons: A Neural-based Approach to Answering Questions about Images, Mateusz Malinowski1 Marcus Rohrbach Mario Fritz Max Planck Institute for Informatics, Saarbrucken, Germany 2UC Berkeley EECS and ICSI, Berkeley, CA, United States, October 1, 2015

[22] [22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, Imagenet large scale visual recognition challenge, IJCV, 2015.

[23] [23] [T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient estimation of word representations in vector space, in ICLR, 2013.

[24] [24] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, DeViSE: A deep visual-semantic embedding model, in NIPS, 2013.

[25] [25] Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks

[26] [26] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4):541551, 1989.

[27] [27] ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering : https://arxiv.org/pdf/1511.05960.pdf

[28] [28] Lin ,Ma Zhengdong Lu Hang Li Learning to Answer Questions From Image Using Convolutional Neural Network, 13 Nov, 2015

[29] [29] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: visual question answering. In ICCV, 2015, M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In ICCV, 2015

[30] [30] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? Dataset and methods for multilingual image question answering. In NIPS, 2015., A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. arXiv preprint arXiv:1606.08390, 2016

[31] [31] S. Hochreiter and J. Schmidhuber, Long shortterm memory, Neural computation, vol. 9, no. 8, pp. 17351780, 1997

[32] [32] Aishwarya Agrawal , Jiasen Lu , Stanislaw Antol , Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh,VQA: Visual Question Answering, 27th Oct 2016

[33] [33] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen*, Anthony Dick, Anton van den Hengel Visual Question Answering: A Survey of Methods and Datasets

[34] [34] Mengye Ren, Ryan Kiros, Richard Zemel Exploring Models and Data for Image Question Answering 8 May 2015

[35] [35]Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, Lucy Vanderwende Generating Natural Questions About an Image , 2016 March
[36] [36]Mengye Ren, Ryan Kiros, Richard Zemel May 2015 Exploring Models and Data for Image Question Answering