A Novel Approach on Visual Question Answering by Parameter Prediction using Faster Region Based Convolutional Neural Network

Sudan Jha [1], Anirban Dey [1], Raghvendra Kumar [2], Vijender Kumar-Solanki [3]*

[1] Kalinga Institute of Industrial Technology, Bhubaneswar (India)
[2] LNCT College, Jabalpur (India)
[3] CMR Institute of Technology, (Autonomous), Hyderabad, TS (India)

Received 16 April 2018 | Accepted 31 July 2018 | Published 31 August 2018

LA UNIVERSIDAD
EN INTERNET

Abstract

Visual Question Answering (VQA) is a stimulating process in the field of Natural Language Processing (NLP) and Computer Vision (CV). In this process machine can find an answer to a natural language question which is related to an image. Question can be open-ended or multiple choice. Datasets of VQA contain mainly three components; questions, images and answers. Researchers overcome the VQA problem with deep learning based architecture that jointly combines both of two networks i.e. Convolution Neural Network (CNN) for visual (image) representation and Recurrent Neural Network (RNN) with Long Short Time Memory (LSTM) for textual (question) representation and trained the combined network end to end to generate the answer. Those models are able to answer the common and simple questions that are directly related to the image's content. But different types of questions need different level of understanding to produce correct answers. To solve this problem, we use faster Region based-CNN (R-CNN) for extracting image features with an extra fully connected layer whose weights are dynamically obtained by LSTMs cell according to the question. We claim in this paper that a single R-CNN architecture can solve the problems related to VQA by modifying weights in the parameter prediction layer. Authors trained the network end to end by Stochastic Gradient Descent (SGD) using pre-trained faster R-CNN and LSTM and tested it on benchmark datasets of VQA.

Keywords

Natural Language Processing, Computer Vision, Convolutional Neural Network, Recurrent Neural Network, Long Short Term Memory, Faster Region Based Convolutional Neural Network, Stochastic Gradient Descent.

DOI: 10.9781/ijimai.2018.08.004

I. Introduction

Understanding an image by the help of computer vision or image processing technique is a complex procedure studied in the two last eras. Then the scientists introduced compatible circumstances between Image Processing and Natural Language Processing (NLP) to solve the problem of image understanding. Traditionally the researchers all over the world applied the process of image understanding to solve the problem of Visual Question Answering by the machine learning. Recently, deep learning architectures constructed by knowledge artificial neural networks have enhanced visual image understanding [1, 2, 3].Object recognition from an image is done by Convolutional Neural Network (CNN). Recurrent Neural Network (RNN) with Long Short Term Memory (LSTM) cell has outstretched the bar on sequence prediction jobs as well as machine translation [18, 19]. CNN performs feature representation of the image and LSTMs process the representation of question and answer. The researchers directly combined both networks and trained end to end to generate the answer [20, 21]. But this kind of approach is able to answer the common and simple questions that are related to the image's content i.e. 'What is the shape …?' or 'How many?' To find a correct answer for the different types of questions, understanding of an image should be different [4, 5].

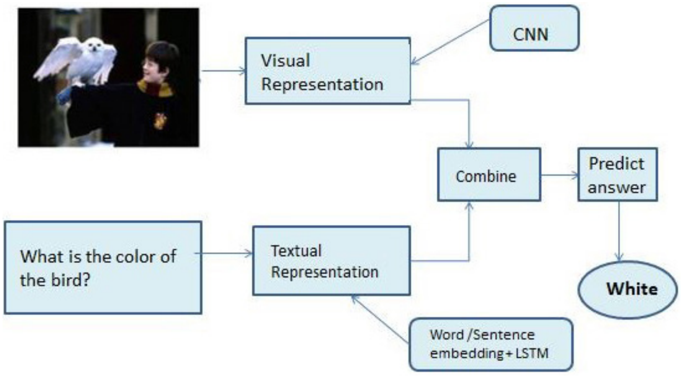


Fig. 1. Visual Question Answering: Common Approach.

Fig.1 indicates the steps of VQA approach at a glance. First image and question need the visual and textual representations. CNN provides the visual representation (feature vector of image) and Word or sentence embedding followed by RNN +LSTM provides textual representation (feature vector of question). Then both features are combined together for predicting the answer [6, 7, 8].

* Corresponding author.
E-mail address: spesinfo@yahoo.com

There are various kinds of questions and answers which are used to train the VQA system. Questions can be mainly of these types: fine grained recognition (e.g. "What kind of animal is in the ground?"), Object detection (e.g. "How many birds are there?"), activity recognition (e.g. is this boy smiling?"), knowledge based reasoning (e.g. "which animal is the reptile in this picture?"), commonsense reasoning (e.g. "Is the boy expecting a job?"). Answer format can be of two categories: yes/no (binary) or multiple-choice. For open-ended, formless questions, answer has not any particular form it can be one word or several words. Images can be mostly two types: Real (real world images like photograph of nature, human, object, etc.) and abstract (clip-arts, animated scene). That is why one type of object recognition for the image is not sufficient for visual question answering problem; selection of the appropriate recognition task is also significant [9, 10].

Computer vision and NLP community released different types of approaches among those techniques, actually the main methodology is to combine the image and question features to create deep learning architecture for classifying jointly combined features to produce a correct answer. Some approaches use CNN for extracting image features and bag of words (BOW), for obtaining question features [11, 12].

VQA is a complete AI finish task and it is multi-disciplinary and heterogeneous in nature. The visual/ image question answering model can solve the complex real life problems (discussed in future work). Another point is that it tackled different sub problems in a single task i.e. image processing, question answering in NLP and sometimes knowledge representation. Motivation from this research problem is a variety of standpoints like new datasets and techniques in computer vision and image processing, integration of vision, language and common sense and merging group of methods of machine learning and deep learning. It has the ability to train the machine how to learn and answer the question like human being and also learn about the pictorial world [13]. The VQA system can be able to learn what do need to learn and what do not need. VQA system is relevant to various types of social applications that include situational information from visual content, making decisions from huge amount of investigation data and, last but not the least, interaction with robots. This research has the purpose to improve the lives of visually-impaired people and transfigure the society through interaction between human beings and visual information [14].

VQA needs various types of understanding of an image not only caption generation or image entity recognition but visual scene understanding and reasoning of knowledge about to the image. The approaches that we discuss in related work took the VQA problem as a classification task, in addition to some network architectures designed to solve the above mentioned problem in joint embedding manner. To overcome the previous problem, we use Faster Region based CNN (R-CNN) for feature extraction of image and updating weights of that CNN according to the question. For weights (parameter) updating we use different types of functions into the weights matrix of the main network. Actually we are trying to generate the answer of free-form and open ended question from the image.

In the section II, we discuss the recent existing approaches. It divided the approaches into four different types; without deep learning based approaches [1, 2], deep leaning based approaches [3, 5, 6, 15, 21], deep learning based approaches with attention [15, 16], other different approaches [17].

In the section III we discuss the basic concepts behind the deep learning tools CNN, RNN, LSTM and Word embedding which are applied to build model to solve VQA problem.

In the section IV & V we describe our proposed model and methodologies to solve the problems over VQA in field of CV and NLP. Our key contributions to this research are that we describe a different methodology that depends on the question part. We have applied faster R-CNN (Region based CNN) [7] with a prediction of parameter layer where the weights are updated according to the question. We apply the word embedding method [12] to question and build a parameter prediction network with a series of LSTM cells. In the section VI we conclude proposed work in details. It uses pre-trained faster R-CNN [7] and LSTM [9] to initialize the weights of entire network by using Stochastic Gradient Descent [22] for training the whole network.

II. Related Work

Malinowski et al. [2] proposed a model that can answer the probability of a given question according to an image. This model gives a single word answer. Authors proposed another model which produces multi word answer. This model is trained and tested on DAQUAR dataset.

Kaffle et. al. [1] proposed a model that is based on Bayes' rule of probability. They trained on COCO-QA data set that contains four types of answers such as counting, object, location; color. The model evaluates the probability of answer dependent on answer type.

Malinowski et al. [3] proposed a method called Neural image QA where they jointly fed image and question features into LSTM encoder-decoder to generate answer. Image features are extracted by CNN and question embedding is obtained from LSTM. The main limitation of the model is that it can be applied on large dataset only. They also propose two metrics i.e. "Average Consensus" which defines human disagreement and "Min Consensus" which defines the disagreement in question answering of a human.

Ren et al. [6] proposed a method that uses RNN and visual semantic embedding in the middle phases of image segmentation and object detection. Their main contribution is the question generation algorithm that transforms image descriptive dataset into question - answer format. They treated the VQA problem as classification rather than answer generation.

Gao et al. [21] proposed a little different procedure that uses LSTMs to translate the question and create the answer by two diverse forms. First LSTM is shared weight mechanism into LSTM encoder –decoder architecture that focuses on grammatical structure of the sentence. Second LSTM extracts features from image by CNN that are not directly served into encoder at each time step.

Noh et al. [5] derived a new approach named "DPPnet" of VQA which learns by CNN with dynamic parameter prediction layer. Here weight matrix of the CNN is dynamically predicted by a parameter predictive network. Parameters were predicted by Gated Recurrent Unit (GRU), which takes the asking, question as an input and produce candidate weights as an output. They used hashing technique for the arrangement of final weight matrix of CNN. This method improved the accuracy on the benchmark data set.

Lin et al. [14] did not utilize any RNN for encoding question. They utilize 3 kinds of CNN for answering the question from an image: CNN for image, CNN for sentence and extra multimodal CNN. In the sentence CNN they utilize 3 layers of convolution and max pooling for creating the sentence portrayal. Catching the relations amongst feature vector of image and question are ended by multimodal convolution step. They added two extra fully connected layer; one for multimodal convolution i.e. they embedded one image feature in between two consecutive semantic components of question side and other one is softmax classification layer to predict answer.

Shih et al. [15] proposed an attention-based model henceforth referred to as Where To Look (WTL). The authors use VGG net to extract the image features .They average the each word vector in the question to obtain the question vector.

Then they compute attention vector over the set of the image features to choose which region to be focused on. The final image representation is the summation of attention weighted regions of the image. After that they concatenated between question features and summed result and fed through full connected softmax layer to predict answer.

Yang et al. [16] proposed a model called Stacked Attention Networks (SAN) encode question using either CNN or LSTM. Then the question encoding is fed to attend over image. Then question encoding and the attention weight are concatenated to calculate attention over the unique image. This model increases the accuracy of VQA system for use of attention.

Andreas et al. [17] proposed a new approach called Neural Memory Network (NMN) that is based on semantic parsing of the question. Parse tree is converted to module to answer the question. Modules are compassable and independent. The question parsing is performed to identify the grammatical relation between parts of the sentence. They use ad-hoc handwritten rules to convert the parse tree in structured queries. The main problem in this model is parsing of a question due to fixed network structure. Error cannot be recoverable.

III. Preliminaries

A. Convolutional Neural Network (CNN)

CNN is used to extract the feature vector from an image. There are two phases in CNN: feature extraction and output prediction. There are two layers in feature extraction: convolution layer and sub sampling (pooling) layer. And one fully connected layer for classification or output prediction. After convolution layer the obtained feature goes through an activation function. In the convolutional layer there are series of matrix multiplications followed by summation operation.

B. Recurrent Neural Network (RNN)

Recurrent Neural Network [4, 24, 25] is used for learning the sequence of data like series of video frame, text, music etc. The main difference of RNN over feed forward ANN is that we add another weight matrix, that matrix comes from previous hidden state. We just give input of the hidden state in every time step, and keep repeating that. Main problem with the RNN is vanishing gradient problem i.e. whenever we do back propagating in the neural network gradient (product of partial derivative of error) tends to vanish. The mathematical formulation RNN is the following:

$$(O)^t = f(h^t \; ; \; \theta_o) \tag{1}$$

$$h^t = f(h^{t-1}; x^t; \; \theta_h) \tag{2}$$

Where o^t is the output of the RNN and x^t is the input at time t, h^t is the state of the hidden layer(s) at time t. θ is encapsulated parameter of weight and bias. Fig. 2 defines a simple graphical model to explain the relation between these three variables in an RNN computation graphs.

In Fig. 2, the values θ_i, θ_h, θ_o represent the parameters associated with the inputs, previous hidden layer states, and outputs, respectively.

C. Long Short Term Memory (LSTM)

LSTM [9] resolves the problem of vanishing gradient and getting successes in natural language processing applications like machine translation, speech recognition etc. LSTM cell consists mainly three gates i.e. input gate, output gate and forget gate and a cell state. Actually LSTM says what is the relevant part of that a network has learned and what to forget.

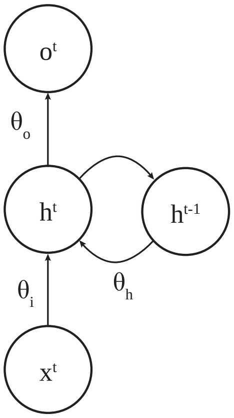


Fig. 2. A graphical model for an RNN.

D. Long Short Term Memory (LSTM)

LSTM [9] resolves the problem of vanishing gradient and getting successes in natural language processing applications like machine translation, speech recognition etc. LSTM cell consists mainly three gates i.e. input gate, output gate and forget gate and a cell state. Actually LSTM says what is the relevant part of that a network has learned and what to forget.

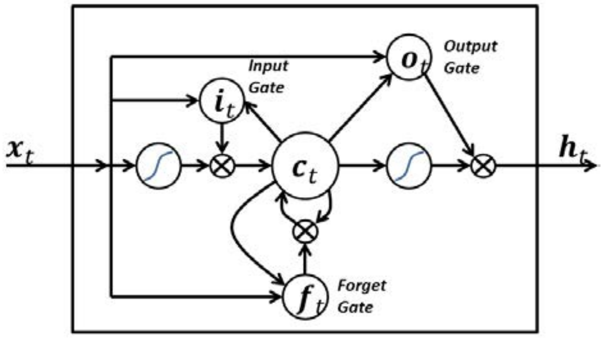


Fig. 3. State diagram of LSTM cell.

In Fig. 3, f_t, i_t, o_t are variables; forget gate, input gate and output gate, respectively. W_f, W_i, W_o are the corresponding weight vectors and b_f, b_i, b_o are the bias. C_{t-1}, h_{t-1} are values of previous cell and previous hidden sate respectively at time t-1.c_t is the value of cell state at time t; h_t is the output vector of LSTM cell at time t.

Cell state is long-term memory, it represents all the learning at over time, and hidden state (is like a current memory). Forget gate is called remember vector that learns what to forget and what to remember. Input gate is also called the save vector that determines number of inputs getting into the cell state, and output gate is also called focus vector that is akin to an attention mechanism (what part of data should be focused on). Actually these gates of LSTM are perception i.e. single layer neural network. So LSTM functionality is mainly based on forgetting, remembering and paying attention of data. The mathematical formulation LSTM is the following:

$$f_t = \sigma(W_f.[C_{t-1}, h_{t-1}, x_t] + b_f) \tag{3}$$

$$i_t = \sigma(W_i.[C_{t-1}, h_{t-1}, x_t] + b_i) \tag{4}$$

$$C_t = Tanh(W_c.[h_{t-1}, x_t] + b_c) \text{ "}Memory\ cell\ state\text{"} \tag{5}$$

$$c_t = f_t.[c_{t-1}] + [i_t].C_t \tag{6}$$

$$o_t = \sigma(W_o.[C_{t-1}, h_{t-1}, x_t] + b_o) \tag{7}$$

$$h_t = o_t.[Tanh\ (C_t)] \tag{8}$$

In the equation (3), the output of the forget gate (f_t) denotes the cell state that what to forget by multiplying 0 to the particular position in the LSTM matrix. The information is remembered if f_t is equal to 1. Here sigmoid activation function σ is applied to the weighted input and previous hidden state.

In the equation (4, 5), the output of input gate (i_t) is a sigmoid function ranged of [0, 1]. That is why sigmoid function is not capable to forget the information of the cell state. The output of input modulation gate (C_t) is *tanh* activation function ranged of [-1, 1]. It permits the cell state to forget the information.

Equation (6) is called cell state equation. The previous cell state c_{t-1} forgets information multiplying by output of forget gate f_t and adds new information through the output of input gate (i_t).

Equation (7) is called output gate equation. The output of the output gate equation (o_t) defines all possible values from LSTM matrix which must be moving forward to the next hidden state.

In the equation (8), h_t is output of hidden state equation that defines what information we should take for next sequence.

E. Word Embedding

Words can be represented as a vector of real valued numbers. In the vector representations words with similar vectors should be semantically the same, that is, they have to represent related concepts. Sometime for practical purposes these vector can have the dimension of 10 or 100 as compared to the vocabulary size which is at least dimension 10 or 1000. Word2vec, GloVe, Skip thought [12] are some methods of word embedding.

IV. Proposed Work

In this section we discuss each tool that we use in the model and how they work together for solving this heterogeneous problem.

A. Image Features

In our method, we use faster R-CNN pretrained model VGG-16 [7] (16 layer CNN in faster R-CNN framework).

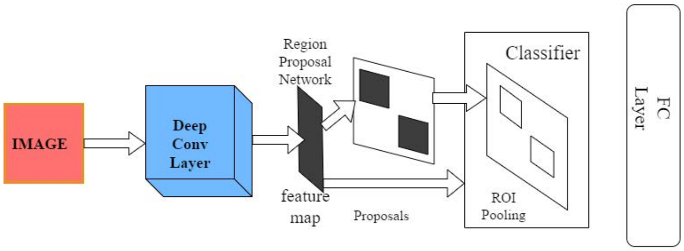


Fig. 4. Faster R-CNN block architecture.

The input image is fed through the Faster R-CNN [7] to get a vector representation. The R-CNN network is trained on Visual Genome [13] dataset for the focusing element of images with the help of image annotation. We vary the value of K (number of image regions) according to the variety of each image. After the convolution layer of faster R-CNN, there is a region proposal network (Fig. 4 and 5) in which some part of the feature pixel matrix of input image is called proposal. Proposal is divided in some regions and selection of maximum feature value from each region is called ROI (Region of Interest) pooling. After the ROI pooling the required image feature vector is generated. We remove the last layer which is fully connected and attached three fully connected layers. The second last fully-connected layer is the parameter prediction layer whose weights are updated according to the question. The dimension of final output vector is equal to the number of possible answers. In the final layer (fully connected) softmax classifier [23] is applied to the output vector for calculating the probabilities of each answer. Output of the classification network is denoted by O = [o_1.......o_n] ^{T} is

$$O = W_r(Q) \times I + b^r \tag{9}$$

Where I is the input vector of parameter prediction layer I = [i_1........i_n]^T, W_r (Q) ∈ $R^{M \times N}$ represents the matrix constructed by LSTM network given to the question Q. b is bias.

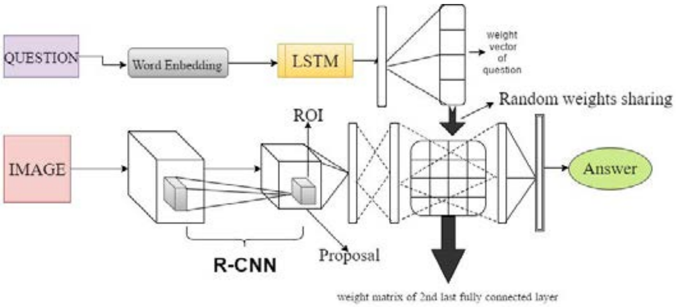


Fig. 5. Network architecture for VQA system.

B. Question Features

First we tokenize the question sentence means splitting the words with space and punctuation and any number is considered as words also. Questions are clipped into 14 words sentence and extra words are thrown out. Then we use pretrained GloVe word embedding (Global Vector for Word Representation). Each word and their corresponding vector have dimension of 300. We also use padding scheme (vectors of zeros) for the question which length is shorter than 14 words. After word embedding we fed the output sequence into LSTM. Output of the LSTM (Q_l) is multiplied by weight vector (W_l) of last fully connected layer of LSTM for getting the candidate weight vector of LSTM network. Candidate weight vector denoted by L= [l_1.........l_z] is

$$L = W_l \cdot Q_l \tag{10}$$

Where W_l is the weight matrix of a fully connected layer of LSTM network and Q_l is final output vector of question.

C. Parameter Prediction

We update the weight (parameter) matrix of 2nd last layer of R-CNN on the basis of the candidate vector of LSTM network. Here we have used parameter reduction techniques to optimize the network. Here the weight matrix of 2nd last fully connected layer of faster R-CNN is denoted W_r (Q) = [w^r_{11}.........w^r_{kl}].

Now, w^r_{kl} is a weight of the weight matrix W_r (Q), it is the corresponding weight between kth output and lth input neuron. £(k, l) is a function to map a key (k, l) to a natural number in {1,..., Z}, where Z is the dimensionality of l. The final value is given by:

$$W_{k-1}{}^{f} = l_{\varepsilon\,(k-1)} \cdot \, \omega\,(k, l) \tag{11}$$

Where ω (k, l) ∈ N × N → {+1, −1} is another function to remove the bias. By these two functions we update the weight matrix W_r (Q) according to the question Q.

Algorithm: *Proposed Algorithm*

```
Input: I , Q   (I is input image , Q is input question)
Output:  A ( A is predicted answer of the question Q about the image I)
1. Extract the image features by FasterR-CNN, output of feature vector
   V size of (k×2048). k = number of image location
2. Resize the feature maps (14×14) by ROI pooling operation into (7×7).
3. Embed the question Q by GloVe word embedding into (14×300)
   vector Q_e.
```

4. *Q_e is fed to the LSTM which internal state dimension is 512 and number of hidden layer MLP is 3.*
5. *After question embedding output is final vector obtained from series of LSTM cell Q_l.*
6. *Now the weight vector of the LSTM network $l=W_l \cdot Q_l$; where W_l is the weight matrix of a fully connected layer of LSTM network and Q_l is final output vector of question.*
7. *The weight matrix of 2nd fully connected layer of R-CNN is denoted by W_r (Q) and we update the W_r (Q) with respect to l (candidate weight vector obtained from LSTM network) by random weight sharing method. In other words, the weight matrix corresponding to the layer is parameterized by a function of the input question Q.*
8. *In the final layer softmax classifier is applied to the output vector for calculating the probabilities of each answer.*

D. Training

Stochastic gradient is used to train the network. Questions are preprocessed by tokenization method [10] and transformed the upper case letter to lower case before training. We do the same thing for answers. To prevent over fitting we identify the epoch which gives the best performance on VQA dataset. Then the training is repeatedly running for the similar number of epochs. Optimization of weight updating layer of R-CNN changes in each batch. For clip gradient of LSTM networks we identify the range of gradient norm and then scale down the exceeding gradient range. We reuse the hyper parameters to normalize the loss. When training time of LSTM goes long then we stop the training because loss of the training does not improve for several epochs. When LSTM network starts to over fit, we stop tuning LSTM but training on the other parts of the network is going on.

E. Training Error by Stochastic Gradient Descent (SGD)

The use of SGD in the neural network is driven by the high cost of running back propagation over the full training set. SGD can overcome this cost and still lead to fast convergence. Back propagation gives you the gradients, but not how to use them. SGD optimizes the gradient and computes the objective function; the standard gradient descent algorithm updates the parameters θ of the objective function J (θ) as,

$$\theta = \theta - \alpha \nabla\theta E \times J(\theta) \tag{12}$$

The above equation (12) is approximated by evaluating the cost and gradient over the full training set. Stochastic Gradient Descent (SGD) simply does away with the expectation in the update and computes the gradient of the parameters using only a single or a few training examples.

V. Experiment

We tested our model in various VQA datasets like COCO-QA, VQA, DAQUAR-all and DAQUAR-reduced [13] and tested before fine-tuning of LSTM and RCNN and after fine-tuning. We also describe the differences between existing methods and our method. And test results in terms of image and question also is displayed in this section.

TABLE I. Empirical Comparison Between Existing Models and Our Model

Models	Methods	Answer Type	Image Features
Neural –Image –QA [3]	Joint embedding	Generation	GoogLeNet
VIS +LSTM [6]	Joint embedding	Classification	VGG-Net
Multimodal–CNN [14]	Joint embedding	Generation	GoogLeNet
LSTM+Attention [20]	Attention method	Classification	VGG-Net
StackAttn network [16]	Attention method	Classification	VGG-Net
iBowing [18]	Joint embedding	Classification	GoogleNet
Bayesian [1]	Joint embedding	Classification	VGG-Net
NMN [17]	Compositional model	Classification	VGG-Net
Attribute LSTM [19]	Knowledge base	Generation	VGG-Net
Proposed Work	**Parameter prediction**	**Classification**	**VGG-Net (Faster R-CNN)**

Table I indicates the differences between previous approaches and our approach on basis of the method used, answer type and image features (pre-trained CNN model name). Joint embedding means combining both image and question representations together. Attention method means focusing the particular regions of the image or questions rather the whole image or question. Knowledge base signifies facts about the world. It has an inference engine for reasoning about the facts by using rules and logic, and deduces the new facts.

A. Datasets

Generally Datasets are prepared of triplets (question, answer, and image). Some number of datasets contains some kind of additional notation. Most of the answers are single words or phrases.

Table II includes the description dataset in terms of number of images, number of questions, number of question types (number questions, color questions, object question and location questions), and number of training and testing questions, question collection and evaluation metrics.

B. Evaluation Metrics

In DAQUAR and COCO-QA dataset, there are two types of evaluation metrics; one is classification accuracy and another one is WUPS (Wu-Palmer similarity [6]) based on WordNet [5] categorization to calculate the semantic similarity between words. It signifies the similarity between machines generated answer and human answer.

Accuracy: The accuracy of classification is calculated by the following equation:

$$Accuracy = \frac{Item\ Classified\ Correctly}{All\ items\ classfied} \tag{13}$$

WUPS (Wu-Palmer similarity): The Wu & Palmer calculates relatedness by considering the depths of the two synsets in the WordNet taxonomies, along with the depth of the LCS (Longest Common Subsequence).

TABLE II. Brief Description About Some Well-Known Dataset of VQA

Dataset	No. of images	No. of questions	No. of question types	Question collection	No. of training questions	No. of testing question	Evaluation metrics
DAQUAR[13]	1449	12468	4	Human	3,876	297	Acc.& WUPS
COCO-QA[13]	117,684	117,684	4	Automatic	78736	38948	Acc. & WUPS
VQA[13]	204,721	614163	20+	Human	248349	244302	Acc. against 10 humans
Visual genome [13]	108,000	1,445,322	7	Human	-	-	Acc.

$$WUPS = \frac{1}{\mathrm{N}} \sum_{i=1}^{n} \mathrm{Min}\left(\prod_{a\epsilon \mathrm{A}^{t}} \mathrm{Max}_{\,t\varepsilon \mathrm{T}^{i}}\ \mu\,(a,t), \prod_{t\epsilon \mathrm{T}^{t}} \mathrm{Max}_{\,t\varepsilon \mathrm{T}^{i}}\ \mu\,(a,t) \right) \tag{14}$$

Where μ (a, t) defines the threshold Wu-Palmer similarity between prediction (a) and ground-truth (t). 0.9 and 0.0 are two threshold values in our evaluation. This means that 0 <= score < 1.

C. Results and Discussion

In this section, results of the proposed model are tested by using four types of datasets i.e. DAQUAR –all, DAQUAR-reduced, COCO-QA, VQA test-dev. It describes the comparative analysis with the existing methods with respect to accuracy and WUPS score.

TABLE III. Result on DAQUAR- all Dataset

Models	Accuracy (%)	WUPS@ 0.0	WUPS@ 0.9
Neural QA[3]	19.43	25.25	62.00
3-CNN[14]	23.40	29.59	62.95
Attribute LSTM[19]	24.27	30.41	62.29
DPPNet[5]	28.98	34.80	67.82
Bayesian[1]	28.96	34.74	67.33
Proposed Work	**29.92**	**35.56**	**72.79**

TABLE IV.Result on DAQUAR- Reduced Dataset

Models	Accuracy (%)	WPUS @ 0.0	WPUS @ 0.9
GUESS	18.24	29.65	77.59
VIS +BOW[6]	34.17	44.99	81.48
VIS+LSTM[6]	34.41	46.05	82.48
Neural QA[3]	34.68	40.76	79.54
3-CNN[14]	39.66	44.86	83.06
Attribute LSTM[19]	40.07	45.43	82.67
DPPNet[5]	44.48	49.56	83.95
Bayesian[1]	45.17	49.74	83.95
Proposed Work	**47.28**	**51.76**	**88.56**

TABLE V. Results on COCO-QA Dataset

Model name	Accuracy (%)	WUPS @0.0	WUPS @0.9
GUESS	6.65	17.42	73.42
VIS +LSTM[6]	53.31	63.91	88.58
3-CNN[14]	54.95	65.36	88.58
VIS+BOW[6]	55.95	66.78	88.99
Attribute LSTM[19]	61.38	71.15	91.58
DPPNet[5]	61.19	70.84	90.61
Bayesian	63.18	73.14	91.32
Proposed Work	**64.71**	**74.87**	**92.56**

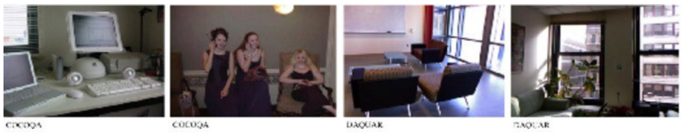
Table III, IV and V show comparison of the results obtained by our proposed model and other existing models "Neural QA, 3-CNN, Attribute LSTM, DPPNet, Bayesian" in terms of both metrics accuracy and WPUS@0.0 and WUPS@0.9 score. The proposed algorithm beats all existing approaches reliably in all benchmarks.

TABLE VI. Result on VQA (test_dev) Dataset

Models	Open ended questions (Accuracy %)	Multiple choice questions (Accuracy %)
Question	43.09	58.68
Image	28.15	30.53
Q+I	52.64	58.97
LSTM Q	48.76	54.75
LSTM Q+I	53.74	57.17
DPPNet [5]	57.22	62.48
Proposed Work	**61.99**	**67.13**

Table VI contains the results of VQA (test-dev) on the open-ended and multiple-choice (M.C.) settings. Here we see that our model performs well in open-ended question as well as in multiple choice questions. The model is tested on only question, only image, question and image both and reported the accuracy.

Now some test images are displayed in Fig. 6 and 7 and shows how the proposed model performs accurately in these different type of questions.



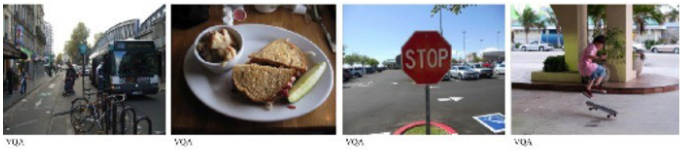
Q: What is the colour of the desk?
Our model: white

Q:what is the colour of the dresses?
Our model: purple

Q: What is the object in front the sofa:
Our model: table

Q: What is the largest Object?
Our model: sofa

Fig. 6. Result of our proposed model in the terms of Question and Image on COCO-QA and DAQUAR Dataset.



Q: How many bikes are there?
Our Model: 2

Q: What is the shape of the plate?
Our Model: Round

Q: What does the sign say:
Our Model: Stop

Q: What type of the tree are here?
Our Model: Palm

Fig. 7. Result of our proposed model in terms of Question and Image on VQA dataset.

The qualitative results of the proposed model are introduced in Fig. 6. As a rule, the proposed method is effective to deal with different sorts of inquiries that need diverse levels of semantic understanding. It demonstrates that the system is able to recognize the task relying upon questions. On the other hand, in Fig.7, the proposed model is compelling to find the answer of different questions on various images.

Some major advantages of our proposed model are discussed in the following:

- Accuracy of this model is better than existing previous VQA models. Because we use single faster RCNN pre-trained model for obtaining visual also.
- We solved the VQA problem not only as a joint embedding approach but we update the weight (parameters) of the network (RCNN), that is why in the last classification layer the number of probabilities (classes) is less than in the other approaches.
- We generate the answer of free-form and open ended question from the image.

CNN entirely loses all their inner data about the position and the placement of the object and they route all the info to the same neurons that may not be able to deal with this kind of information. A CNN makes estimations by looking at an image and then testing to see if assured modules are present in that image or not. If they are, then it classifies that image accordingly. Fine tuning of hyperparamters of our proposed model is non-trivial and we need a large dataset for training the model. We identify that when we trained our model in DAQUAR reduced dataset loss after several epochs are increased drastically. These are the complexities of using a CNN architecture.

VI. Conclusion

This paper focuses on a deep learning based model for both open ended and multiple choice questions. It describes several types of experiments and the contributions of each design. It also provides the importance of several mechanisms of a VQA model. It shows how Faster R-CNN magnifies the ability of object detection in an image fastly, increasing overall accuracy by 13% more than other image recognition models "Neural QA, 3-CNN, Attribute LSTM, Bayesian" do. The paper describes how our model updates the weight matrix of fully connected layer of faster R-CNN according to candidate weights of questions. It has been shown that VQA problem can be solved by a single R-CNN model. The following further directions should be investigated by using capsule Network for the object detection part of VQA system.

References

[1] Kafle, K., and Kanan, C. (2016) Answer-type prediction for visual question answering, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4976–4984.

[2] Malinowski, M., and Fritz, M. (2014) A multi-world approach to question answering about real-world scenes based on uncertain input, In Advances in Neural Information Processing Systems, 1682–1690.

[3] Malinowski, M., Rohrbach, M., and Fritz, M. (2015) Ask your neurons: A neural based approach to answering questions about images, In Proceedings of the IEEE international conference on computer vision, 1–9.

[4] Mikolov, T., Karafi´at, M., Burget, L., Cernock`y, J., and Khudanpur, S. (2010) Recurrent neural network based language model. In Inter speech, 3-12.

[5] Noh, H., HongsuckSeo, P., and Han, B. (2016) Image question answering using convolutional neural network with dynamic parameter prediction, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 30–38.

[6] Ren, M., Kiros, R., and Zemel, R. (2015) Exploring models and data for image question answering, In Advances in neural information processing systems, 2953–2961.

[7] Ren, S., He, K., Girshick, R., and Sun, J. (2015) Faster r-cnn: Towards real-time object detection with region proposal networks, In Advances in neural information processing systems, 91–99.

[8] Sabour, S., Frosst, N., and Hinton, G. E. (2017) Dynamic routing between capsules, In Advances in Neural Information Processing Systems, 3857–3867.

[9] Sundermeyer, M., Schlüter, R., and Ney, H. (2012) LSTM neural networks for language modeling, In Thirteenth Annual Conference of the International Speech Communication Association, 147-156.

[10] Teney, D., Anderson, P., He, X., and Hengel, (2017) A. v. d. Tips and tricks for visual question answering: Learnings from the 2017 challenge, 24-31.

[11] Yang, Z., He, X., Gao, J., Deng, L., and Smola, A. (2016) Stacked attention networks for image question answering, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 21–29.

[12] Zou, W. Y., Socher, R., Cer, D., and Manning, C. D. (2013) Bilingual word embeddings for phrase-based machine translation, In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 1393–1398.

[13] Antol, S., A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh (2015) VQA: visual question answering. In ICCV, 24-35.

[14] Ma, Lin, Zhengdong Lu, and Hang Li. (2016) Learning to Answer Questions from Image Using Convolutional Neural Network. AAAI. 3(7), 12-21.

[15] Shih, K. J., Singh, S., & Hoiem, D. (2016). Where to look: Focus regions for visual question answering, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4613-4621).

[16] 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).

[17] Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016).Neural module networks, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 39-48).

[18] Wang, P., Wu, Q., Shen, C., & Hengel, A. V. D. (2016). The VQA-Machine: Learning How to Use Existing Vision Algorithms to Answer New Questions. ArXiv preprint arXiv: 1612.05386, 147-154.

[19] Yao, T., Pan, Y., Li, Y., Qiu, Z., & Mei, T. (2016). Boosting image captioning with attributes. arXiv preprint arXiv:1611.01646, 21-31.

[20] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention, In International Conference on Machine Learning (pp. 2048-2057).

[21] GAO, H., Mao, J., Zhou, J., Huang, Z., Wang, L., &Xu, W. (2015). Are you talking to a machine? Dataset and methods for multilingual image question, In Advances in Neural Information Processing Systems (pp. 2296-2304).

[22] Acharya, U. R., Fujita, H., Lih, O. S., Hagiwara, Y., Tan, J. H., & Adam, M. (2017). Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network. Information sciences, 405, 81-90.

[23] Liu, Y., Chen, X., Peng, H., & Wang, Z. (2017). Multi-focus image fusion with a deep convolutional neural network. Information Fusion, 36, 191-207.

[24] Mansor, M. A., Kasihmuddin, M. S. M., & Sathasivam, S. (2016). Enhanced Hopfield network for pattern satisfiability optimization. International Journal of Intelligent Systems and Applications, 8(11), 27.

[25] Mansor, M. A. B., Kasihmuddin, M. S. B. M., & Sathasivam, S. (2017). Robust Artificial Immune System in the Hopfield network for Maximum k-Satisfiability. International Journal of Interactive Multimedia and Artificial Intelligence 4(4), 63-71.

Sudan Jha

Sudan Jha holds the Ph.D. in CSE , working as professor in School of Computer Engineering, KIIT University, India with 18 years of experience in Academics, Administration as a Principal in four Colleges in India and Nepal, Industrial Research and Software Project Development &Management; As a team leader executed three government funded projects; Board of directors and Technical Advisor in Nepal's National Television (NTV); Chief IT Consultant in Nepal Telecom Authority – the regulating body of telecommunication of Nepal; published so far 31 International Journal transaction papers (SCOPUS Indexed-3, UGC Approved-8, Others-20); 60+ conference papers; 8 major software developed as Team Leader including Govt and non Govt. organization. Editorial board member in several Journals, International Conferences, Books. Provisional issuance is under process for his second Ph.D. in Computer Engineering which is already under the final verge. He has been an invitee for ROBOTRONICSUSA twice, World Youth Festival Russia as a VIP Guest Speaker, Teacher's Education Club, Harvard University USA as an active audience. He has visited 14 countries and delivered talks in various topics related with Network Systems, Embedded Systems, Parallel & Distributed Systems, Operating System, Distributed Operating System, Programming Skills like Programming in C, Programming in C++, Object Oriented Software Engineering, Emerging Subjects like High Speed Downlink Packet Access Networks (HSDPA), Universal Mobile Universal Telecommunication Systems (UMTS), Cell throughput evaluation of HSDPA Networks, Fundamentals of Software Engineering, Data Mining and Warehousing, Wired and wireless networking, Software Engineering, Data Mining and Warehousing, UMTS based HSDPA. He also holds standing on consultant for the Nepal Telecommunication Authority, the regulatory board for Government of Nepal and member of board director of Nepal Television.

Anirban Dey

Anirban Dey pursuing M.Tech in computer science and engineering from KIIT University and a research scholar. He is doing researches in the field of computer vision, Natural Language Processing and Deep Learning. His some papers were published in IEEE-approved and Springer conference.

Raghvendra Kumar

Raghvendra Kumar is working as Assistant Professor in Computer Science and Engineering Department at L.N.C.T Group of College Jabalpur, M.P. India. He received B. Tech. in Computer Science and Engineering from SRM University Chennai (Tamil Nadu), India, M. Tech.in Computer Science and Engineering from KIIT University, Bhubaneswar, (Odisha) India and Ph.D. in Computer Science and Engineering from Jodhpur National University, Jodhpur(Rajasthan), India. He has published 86 research papers in international / National journal and conferences including IEEE, Springer and ACM as well as serve as session chair, Co-chair, Technical program Committee members in many international and national conferences and serve as guest editors in many special issues from reputed journals (Indexed By: Scopus, ESCI).He also received best paper award in IEEE Conference 2013 and Young Achiever Award-2016by IEAE Association for his research work in the field of distributed database. His research areas are Computer Networks, Data Mining, cloud computing and Secure Multiparty Computations, Theory of Computer Science and Design of Algorithms. He authored 12computer science books in field of Data Mining, Robotics, Graph Theory, and Turing Machine by IGI Global Publication, USA, IOS Press Netherland, Lambert Publication, Scholar Press, S. Chand Publication and Laxmi Publication.

Vijender Kumar-Solanki

Vijender Kumar Solanki, Ph.D. is an Associate Professor in Computer Science & Engineering, CMR Institute of Technology (Autonomous), Hyderabad, TS, India. He has more than 10 years of academic experience in network security, IoT, Big Data, Smart City and IT. Prior to his current role, he was associated with Apeejay Institute of Technology, Greater Noida, UP, KSRCE(Autonomous) Institution, Tamilnadu, India & Institute of Technology & Science, Ghaziabad, UP, India. He has attended an orientation program at UGC-Academic Staff College, University of Kerala, Thiruvananthapuram, Kerala & Refresher course at Indian Institute of Information Technology, Allahabad, UP, India. He has authored or co-authored more than 25 research articles that are published in journals, books and conference proceedings. He has edited or co-edited 4 books in the area of Information Technology. He teaches graduate & post graduate level courses in IT. He received Ph.D in Computer Science and Engineering from Anna University, Chennai, India in 2017 and ME, MCA from Maharishi Dayanand University, Rohtak, Haryana, India in 2007 and 2004, respectively and a bachelor's degree in Science from JLN Government College, Faridabad Haryana, India in 2001. He is Editor in International Journal of Machine Learning and Networked Collaborative Engineering (IJMLNCE) ISSN 2581-3242, Associate Editor in International Journal of Information Retrieval Research (IJIRR), IGI-GLOBAL, USA, ISSN: 2155-6377 | E-ISSN: 2155-6385. He is guest editor with IGI-Global, USA, InderScience & Many more publishers. He can be contacted spesinfo@yahoo.com.