Deep Learning Approach to Visual Question Answering

Team 31

Suraj Kiran Raman
surajkra@umich.edu
University of Michigan
Ann Arbor - MI

Vijayakrishna Naganoor
vijaykn@umich.edu
University of Michigan
Ann Arbor - MI

Shanthakumar Venkatraman
shanthav@umich.edu
University of Michigan
Ann Arbor - MI

Dhruv Agnihotri
dagni@umich.edu
University of Michigan
Ann Arbor - MI

***Abstract*—Answering questions based on real world image is a challenging task that has emerged as an active research topic in the recent years. This problem marks the intersection of Natural language processing and Computer Vision. The challenge lies in not only incorporating the techniques that are best for the image processing and natural language processing but integrating them efficiently for the task of Visual Question answering and also how to best utilize the features from the image and questions. Since most of the current approaches are trained on 2D images they are not robust in dealing with questions that require 3D representation of the scene. Our approach D-VQA and DS-VQA provides a novel way of utilizing the spatial features that increases the performance compared to the earlier methodologies. Our model D-VQA achieves a WUPS(0.0) score of 81.57% on the reduced DAQUAR dataset, which is a significant improvement. We also provide additional insights into the problem by analyzing the results.**

***Keywords*—Visual Question Answering, features, D-VQA, DS-VQA, RNN, LSTM.**

I. Introduction

As vision processing techniques like object recognition and segmentation are improving, there has been an increasing interest in full scene understanding. This understanding can be quantified by answering the questions related to an image. Given a reference image and a question, the task of Visual Question Answering (VQA) is to predict an answer to the question which is related to the image. This task is complex as questions and images are in their own feature spaces and we need to link those features in a cohesive manner to perform this task. So in order to successfully perform the task we need to make sure that the system learns how to answer the question based on the image in a way that mimics human performance.

Question answering has been an active research area for a long time. There have been methods like [1],[2],[3] for finding patterns in the questions and utilizing the patterns for answering the questions. Also there has been ample research on learning the patterns as done in [4].

In recent years Neural Networks have become increasingly complex and have been put to use in multifarious domains owing to its ability to produce impeccable results. So the natural inclination was to make use of neural networks in an efficient way for learning. For the task of Question answering, Neural networks have become very popular. Methods proposed in [5], [6] have made use of Recurrent Neural Networks (RNN) in an efficient manner to answer questions just by learning question answer pairs and without using any visual context.

Fig. 1: Result of the proposed model D-VQA, in comparison with that of Neural-Image QA [24]

Equally strong progress has been made in developing Neural Network techniques for image segmentation and Object recognition. [7] , [8] have shown the success of employing Convolution Neural Networks (CNN) for visual classification problems. Both of these independent blocks which are Neural network for images i.e CNN and Neural Network for Questions i.e RNN must be integrated in a cohesive manner for performing VQA.

Recently [9] have proposed a method in which they have merged both the textual and visual Neural Networks by combining CNN and Long Short Term Memory (LSTM) into an end-end architecture that performs Visual Question Answering.

While [9] is a promising result, when analyzing it, it was observed that there were specific question types whose performance was not good. A significant category of questions were based on some description involving depth in them. So this motivated our work into creating two new models namely Depth-VQA (D-VQA) and Depth-Switch-VQA (DS-VQA) which perform equally well in questions pertaining to Depth in the image by using another CNN. The image is analyzed by CNNs to get spatial and depth features and the question is fed into a LSTM network along with these features.

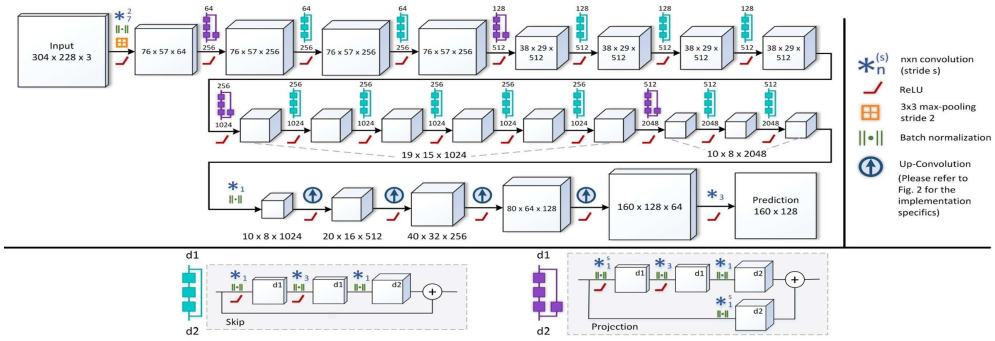

Fig. 2: Architecture for the ResNet-50 Model used to obtain the Depth map of RGB Images

This system is then trained with ground truth and then tested. In this work our main contributions are-:

- We try to exploit the importance of spatial representation of 2d images with respect to VQA with the help of depth features learnt through a new CNN which will be referred as D-CNN.
- We propose two novel methods (D-VQA and DS-VQA) on utilizing both the depth features and global representation features. We combine these features with Questions and pass it to RNN model for training.
- We also study and analyse the performance of multiple optimizers for tuning the parameter of RNN (LSTM) Model.

II. Related Work

A. Global Image representation using Convolutional Neural Network

Deep Convolutional Neural Networks have been successful in addressing the object classification task by virtue of its ability to learn rich-mid level image representations as shown in[7], [8], [10]. The models are learnt on raw image data and are trained on large datasets. There are many complex models that have been developed in recent years [11],[12]. We make use of these models in order to get our features from the images.

B. Sequence Modelling using Recurrent Neural Networks

Recurrent Neural Networks are used to process the sequential data and are generally used in problems of speech processing and language processing like [6],[13] Since it has been shown that Long Short Term Memory (LSTM) can handle the problem of vanishing gradients in an efficient way and give good results for the problems involving Language Processing, [14] we use this method to model our questions.

C. Integration of RNN and CNN for Visual Question Answering

As we are trying to tackle the multi-modal problem of answering a question by looking at the image, we need efficient ways of integrating both the types of inputs. Visual Scene Description has been addressed in [15], [16] to an extent.[17] proposed multi-world approach that conducts the semantic parsing of question and segmentation of image to produce the answer. Deep neural networks are also employed for the image QA task, which is more related to our research work.

In the work of [18], the image QA task is formulated as a classification problem, and the so-called visual semantic embedding (VSE) model is proposed. LSTM is employed to jointly model the image and question by treating the image as an independent word, and appending it to the question at the beginning or ending position. As such, the joint representation of image and question is learned, which is further used for classification. However, simply treating the image as an individual word cannot help effectively exploit the complicated relations between the image and question. Thus, the accuracy of the answer prediction may not be ensured.

Recently in the work of [9] which proposes a system called the Neural Image-QA, the question along with the image is passed through an RNN model consisting of a series of LSTM cells which is later trained to produce the correct answer for that question about the image. The work by [9], formulates the image QA task as a generation problem. [9] will be the main inspiration for our work. We implement the Neural Image-QA system proposed by [9] and we develop our model on top of it.

D. Depth Feature Extraction

As lot of questions related to an image requires the system to make decisions based on the depth, just the global representations are not sufficient to make the decision 5. In that scenario features representing depth must be extracted and integrated with the global features for training. [19] proposes a Residual Convolutional Neural Network model called ResNet-50 whose architecture is shown in 2 which estimates the depth map of a scene given a single RGB image. We use a trained ResNet-50 to obtain the depth features. In particular, we extract the output at fc1000 layer (the one before softmax) once we pass the image through it.

III. Dataset

The experiment can be performed well if we have images, depth map and corresponding questions and answers.

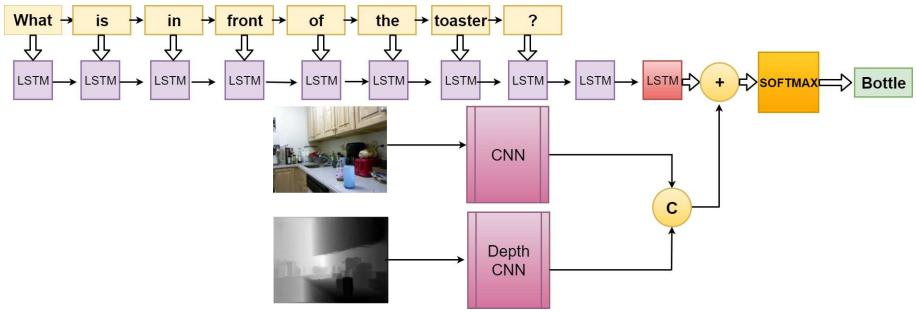

Fig. 3: Visualization of proposed D-VQA model. Each word is passed on through the LSTM cells sequentially and fused with the image representation obtained through CNN and Depth-CNN

In that regard, we make use of the NYU-Depth Dataset V2 [21] which has 1449 images and their corresponding Depth images acquired using Microsoft Kinect. All of these images are collected in an indoor environment. For the question answer pairs corresponding to these images we use a subset of DAQUAR dataset [17] which contains question answer pairs for these 1449 images. DAQUAR dataset on an average consists of 5 question-answer pairs per image in training. We also use another dataset which is a subset of DAQUAR known as Reduced DAQUAR that contains 37 classes and 25 test images. DAQUAR also consists of the comments on challenges associated with every corresponding question-answer which helps us in analyzing the results.

IV. Approach

We have to formulate this problem in a probabilistic way to tackle it efficiently. Given a question and the image, the problem statement at hand can be simplified to predicting an answer in the set of all possible different answers. This can be simplified as estimating probabilities of each word in the set being the answer and choosing that word that corresponds to maximum probability as the answer to this question. Mathematically we can represent this as:

$$A = \arg\max_{a \in B} p(a/q, I; y) \tag{1}$$

where A is the predicted answer and a is the variable over which we maximize the probability which denotes an answer,B is the set of all answers, I represents the image features, q denotes the question features and y represents the vector of all parameters to learn. We later discuss on how to obtain the y,a,q and p(a/q,I;y).

As visualized in 3, for training, the question **X** for a particular image can be treated as a sequence [x_1,x_2,.....,x_{n-1}] of words. The length of this sequence can be estimated by the presence of the question mark at the end of the question. As the sequence **X** can have variable size, we use LSTM to model these sequences.

Now we will describe the working of LSTM cells in detail. The working of the LSTM can be visualized through the figure 4. LSTM's (Long Short Term Memory) have a chain like structure with each repeating unit taking the input sequentially. If x_I is the image representation and x_{qI} is the representation of the question at that time then X_t, the input that is fed into the LSTM cell is given by 2.

$$X_t = [x_I, x_{qI}] \tag{2}$$

Each LSTM cell has an output h(t) which is fed to the

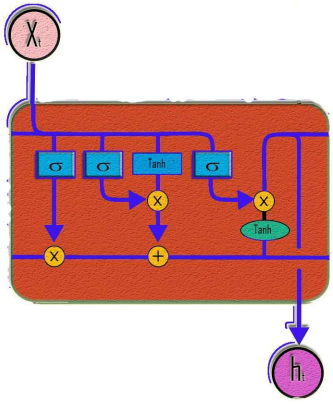

Fig. 4: Overview of architecture of LSTM Cell

next LSTM cell along with the next word in the sequence. The cells have a unit named memory cell that is altered in each of the LSTM cells. It is a crucial component in helping LSTM address the problem of Long Term Dependencies. Every LSTM cell consists of four layers which are denoted in equations with subscripts f, i, C, o interacting in a unique fashion with inputs and themselves. Based on the output of the previous cell $h(t-1)$ and present input $X(t)$, the four layers suggest what part of the information is carried forward in the memory cell and what the new value of the memory cell should be according to equations 3,4,5,6,7,8.

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

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

$$\tilde{C}_t = tanh(W_C.[h_{t-1}, x_t] + b_C) \tag{5}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{6}$$

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

$$h_t = o_t * tanh(C_t) \tag{8}$$

Here C_{t-1} is the old cell state, C_t is new cell state, X_t is the current input, h_{t-1} is output of the previous cell and h_t is output of the current cell. The 8 parameters, four for W and b each, are learnt through training.

For passing the words into the LSTM we have to first convert the words into numerical inputs. Based on all the words in the questions and the answers in the training data, we create an extensive dictionary of words by assigning a unique number for each of the words. Using this dictionary, every word in the question as well as the answer is encoded as one-hot vectors. It is a binary vector which is of the same length as the number of words in the dictionary with the position that has value '**1**' indicating the index of the word in the vocabulary.

We have to pass a multimodal input which consists of the information regarding the image as well as that of the corresponding question through the LSTM. For this purpose, we pass in the image through a ResNet-152, which is a Convolution neural network pre-trained on the ImageNet dataset to get global representation features. The images that we would be dealing with are mainly in the indoor setting which are very close to the domain of the data used in the ImageNet dataset. Hence using the ResNet-152 as a feature-extractor by extracting the output of one of the fully connected layer will give features that represent the image in a nice manner since it is one of the best models available for classification. Assumption is that this output can be treated as 'off the shelf features' for representing the image [22]. ResNet-152 is the best existing variation among the ResNet models [20]. We use the first fully connected layer after the convolution which is a vector of length 2048. This is integrated with the numerical representation of every word and input to the LSTM.

It can be noticed that many questions for these images require the knowledge of the depth as seen in Figure 5. As the global image representations do not capture the idea of depth, we have to come up with better models to represent the depth in an image.

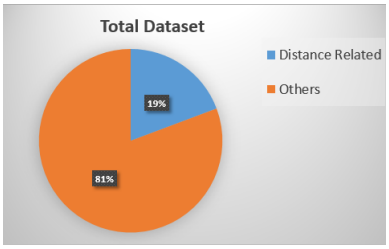

Fig. 5: Percentage of Questions with Depth related words

In order to obtain the depth features for an image we pass the image through the ResNet-50 model shown in Figure 2 and extract the output at fc1000 layer as previously mentioned. This gives us the depth features. We integrate these features along with the global image representations and use it along with the questions. We train two LSTM models, one of them is trained along with the depth features and the the other one is trained using only the global visual representations. This integrated system of LSTM's that we propose is called D-VQA.

However, in questions that do not involve the depth, it might not be necessary to extract and utilise the depth related features. As we have two LSTMs at our disposal, in order to use it efficiently with the above statement in mind, we propose a method called DS-VQA. It involves extracting the depth related features only if the question being asked contains the words that refers to the depth in image. This is decided by comparing the question with dictionary containing depth keywords. This proposed method has been explained in the block diagram shown in Figure 7.

V. Experiments and Results

In this section we evaluate, study and analyze the importance of our proposed D-VQA and DS-VQA model compared to the existing state-of-the art techniques for visual question answering. We also experimented with the parameters used to train the LSTM through validation. In section 5.1, we analyze the performance of various optimizers in learning the LSTM weights. In section 5.2, we experiment on the effect of fusing the multiple feature modalities in predicting the answers. In section 5.3, we discuss our proposed method involving specific attention towards depth representation and benchmark the performance of our D-VQA model. In section 5.4, we compare the refined D-VQA model, DS-VQA on a sampled DAQUAR dataset to study the impact of depth specific switching model on the indoor image visual question answering setting.

Sometimes the answer predicted by the system can be very close to the expected answer but may not exactly match it. If we use a conventional way of calculating the performance of the system, we would be penalizing the system equally for predicting a completely irrelevant answer and answer that is close enough. The metric that is a generalization of the accuracy measure that accounts for word-level ambiguities in the answer has been proposed in [25]. It makes it an ideal metric for reporting accuracy for the problems related to Visual Question Answering. We will be using this metric to report our results. It can be calculated as follows:

$$WUPS(A,T) = \frac{1}{N}\sum_{i=1}^{N} min(\prod_{a\in A^i}\max_{t\in T^i}\mu(a,t), \prod_{t\in T^i}\max_{a\in a^i}\mu(a,t)) \tag{9}$$

where μ is a threshold for measuring the similarity. The metric penalizes more as the threshold increases.

To set up the experiment, we made use of computational packages from Theano with keras wrapper. We have also made use of python based library set developed by Malinowski *et al* [9]

A. Optimizer characteristics

The choice of an optimizer in training a neural network plays a significant role in framing the learned model. In this domain of Deep Neural Nets, optimizers such as Adam, Adamax, SGD, RMSProp, Adadelta etc., are highly prevalent. We experiment the performance of all the above optimizers

with our D-VQA model optimizing the categorical cross entropy loss given by,

$$L_i = \sum_j t_{i,j} log(p_{i,j}) \tag{10}$$

where 't' denotes the target and 'p' denotes the corresponding prediction. Table I depicts the performance of our model across various optimizers mentioned above. Figure 6 depicts the optimizer characteristics (loss vs iterations). As the Figure

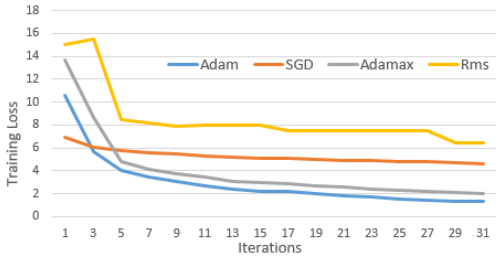

Fig. 6: Performance of optimizers over iterations

6 depicts, the Adam optimizer performs well under this setting of minimizing the categorical cross entropy loss for predicting answers. The Adaptive Moment Estimation (Adam) computes adaptive learning rates for each parameter as opposed to a single learning rate for all the weights in the classical stochastic gradient technique. It combines the advantages of both Adamax and RMSProp in storing an exponential average of the past squared gradients v_t, as well as an exponential decaying average of past gradients m_t, similar to momentum as equations 11,12.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{11}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{12}$$

where β_1, β_2 are the corresponding decay rates. In order to counteract the biases encountered during the initial stages of gradient descent, the algorithm estimates bias corrected moments as equations 13,14.

$$\hat{m_t} = \frac{m_t}{1 - \beta_1^t} \tag{13}$$

$$\hat{v_t} = \frac{v_t}{1 - \beta_2^t} \tag{14}$$

With these bias corrected update parameters, the gradient descent rule is characterized by equation 15

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v_t}} + \epsilon} \hat{m_t} \tag{15}$$

Table I depicts the WUPS scores and convergence characteristics of different optimizers on this Visual Question Answering problem using our proposed D-VQA model. Adam optimizer in our experiment setting, ensures faster convergence than other competing optimizers and hence throughout this work, we use Adam optimizer in all the following experiments.

Models	Accuracy	WUPS@0.9	WUPS@0.0	No. of Iterations
Adam	20.765	27.100	64.110	19
Adamax	18.220	24.916	62.660	30
RMSprop	9.230	20.816	35.710	40+
SGD	4.336	9.127	43.882	40+

TABLE I: Accuracy and WUPS scores for different Optimizers

B. Fusing Features

The proposed D-VQA model, combines three different feature modalities i.e, textual features (questions), visual features (images) and depth features (images). The training set contains 6795 samples with a 10% validation. It is highly important to effectively represent this multimodal feature-set to avoid overfitting in this sample sparse environment. The global visual features and depth visual features obtained through the CNNs are concatenated to form a new set of visual features. We have experimented by fusing the visual and textual features using the sum, concat, average and multiplicative modalities. Table II depicts the WUPS scores obtained when training using the above mentioned modalities on the D-VQA model. Concatenation of visual and textual features yields lesser efficiency because of the Curse Of Dimensionality which makes the system require more training samples to maintain same performance if the dimension of features are increased. Multiplication of these feature modalities does not yield good results, as during multiplication, adverse effect of features over each other reduces the effectiveness of a particular feature in training the system. Summing up the features ensures good performance because there is no adverse effect occurring during summation and we also preserve the original dimensionality of the features. The WUPS scores reported in the table II agree with the intuition stated above. Thus, in the final D-VQA model, we use this summed feature set along with the Adam optimizer to reduce the categorical cross entropy for effectively performing the task of visual question answering.

Model	WUPS@0.9	WUPS@0.0
Sum	27.1	64
Concatination	26.0	62
Multiply	26.8	62.9
Average	24.5	60.2

TABLE II: WUPS scores across various Modalities

C. Evaluation of D-VQA

The main contribution of this work is to ensure that the VQA models specifically learn the depth parameters of an image to ensure improved performances on depth related questions, which form a major part in the indoor DAQUAR dataset developed on the NYU-Depth Dataset V2. Depth features are learnt through an additional CNN as shown in the Figure 3. Table III shows the results of our D-VQA model on the full DAQUAR dataset containing 5673 questions. Amongst these set of questions, a total of 20% questions are related to depth as shown in Figure 5. In comparison with the previous works on visual question answering (Table III), our D-VQA model performs better than [17] with increased accuracy and WUPS scores. This can be mainly inferred as the improvement due to the efficient integration of both visual and textual features. D-VQA performs better than [9] with respect to both test accuracy and the WUPS scores. This improved performance can be attributed to the inclusion of the depth specific learning as in D-VQA, especially given the fact that depth specific questions cover a significant 20% of the entire dataset. We also

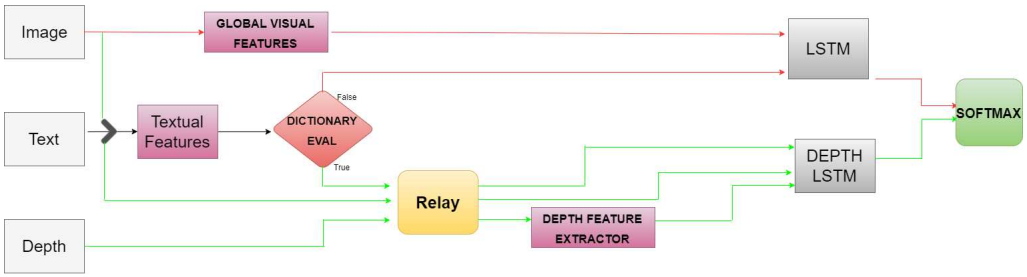

Fig. 7: Visualization of DS-VQA Model. The green line shows the path switched by the model in the event of a depth question

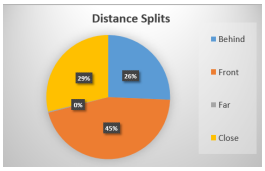

Fig. 8: Depth Question set Composition

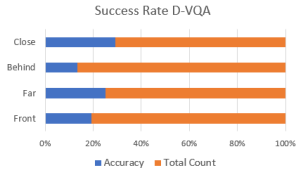

Fig. 9: Success Rate of D-VQA across Depth subclasses

analyze the effectiveness of answering among different kinds of depth specific questions. The composition of depth related question set based on depth keywords is given by Figure 8. The performance of D-VQA across these depth sub-classes is shown in Figure 9. The values reported in Figure 9 are the testing accuracies and we get a much higher WUPS score which is the more relevant metric for the problem of Visual Question Answering. Recently, Reduced DAQUAR dataset

Model	Accuracy	WUPS@0.9	WUPS@0.0
Ours D+N	**20.765**	**27.1**	**64.1**
Multi-World[17]	7.86	11.86	38.79
Ask Your Neurons[9]	19.43	25.28	62.00

TABLE III: Accuracy and WUPS scores of different Models.

which is a subset of DAQUAR dataset that consists of 37 classes and 25 images has garnered attention as the standard dataset for VQA. So we have also performed VQA on this Reduced DAQUAR dataset and benchmarked our performance with the state-of-the-art VQA models whose results are shown in Table IV. As shown in Table IV, we compare the result of D-VQA with existing VQA models. [17] performs the task of answering using a multi-world approach of semantic parsing and segmentation. [23] uses Bag-of-Words features to generate the dense question embedding. The LSTM model answers with just the dense question embedding and hence is commonly termed as the "blind" model. [24] uses attention specific CNNs to improve the visual features for VQA. Human performance for this task of Visual Question Answering is a standard WUPS of 78.96%, Our implementation of [9] achieves a WUPS of 79.54%. Our model D-VQA achieves a WUPS score of 81.57% improving the performances of the existing techniques. Some of the results we obtain for image-question pairs using our D-VQA model are shown in 10,11,12,13

Models	Accuracy	WUPS@0.9	WUPS@0.0
Multi-World[17]	12.73	18.10	51.47
BOW	32.67	43.19	81.30
LSTM	32.73	43.50	81.63
Image+BOW[23]	34.17	44.99	81.43
Neural Image-VQA[9]	34.68	40.76	79.54
D-VQA	**38.72**	**44.38**	**81.57**
ABC-CNN[24]	42.76	47.63	83.04

TABLE IV: Accuracy and WUPS scores of different Models on Reduced DAQUAR Dataset.

D. Evaluation of DS-VQA

Even though D-VQA improves the WUPS score, utilisation of depth features perpetually may lead to reduction in performance on the non-depth question set 12. This inspired us to improve the existing D-VQA model to DS-VQA (Depth Switched VQA) model 7. In this model, we form a dictionary containing the depth specific keywords as in 8 and parse through the input question to determine the existence of any of these keywords. In the event that there is a depth keyword present in the question we make use of the D-VQA model for VQA. If no depth keyword is present we bypass the Depth-CNN (ResNet-50) and use only the Global spatial Image features (ResNet-152) and the Question features. This process effectively creates a switch between the D-VQA model and the "ResNet-152 + LSTM" model.

Model	Accuracy	WUPS@0.9	WUPS@0.0
DS-VQA	**33.2**	**38.2**	**66.83**
D-VQA	29.2	35.03	66.1

TABLE V: Accuracy and WUPS scores of D-VQA and DS-VQA on a sampled DAQUAR dataset

We compare the performance of D-VQA and DS-VQA in a sampled DAQUAR data set. We use a sampled data-set to benchmark our performance due to limitations in computational resources. Table V compares D-VQA and DS-VQA. DS-VQA outperforms D-VQA with a significant 4% increase in accuracy and WUPS scores. This validates our idea on circumstantially utilizing depth features depending on input questions.

What is in front of the toaster?

Our Answer	Neural Image-QA	Ground Truth
bottle	glass	bottle

What is the object on the floor in front of the screen?

Our Answer	Neural Image-QA	Ground Truth
toilet	bathtub	toilet

What is the object in the front?

Our Answer	Neural Image-QA	Ground Truth
table	chair	table

How many tiers does the spice rack have?

Our Answer	Neural Image-QA	Ground Truth
3	2	3

Fig. 10: Image Question pairs where our model gives correct answer and Neural Image-QA model fails.In the case of the first three images illustrated in this figure, the word 'in front' refers to the depth in the image.In such cases our model performs significantly better than Neural Image QA [2]

VI. Conclusion

In this work we have proposed and validated two novel models D-VQA and DS-VQA. We have experimented and concluded that the ADAM optimizer optimizes efficiently in the given VQA setting. We also found that the integration of textual and visual features through summation outperforms the other modalities. With improved accuracy and WUPS scores in both Full and reduced DAQUAR dataset our proposed model D-VQA is highly promising. We have shown that DS-VQA outperforms D-VQA in the constrained environment setting and we expect it to outperform D-VQA significantly with full DAQUAR dataset. In addition to the models we have implemented, we also like to modify our DS-VQA model to include attention based CNN [24] instead of ResNet-152. This can result in the improvement of the model in near future.

What is behind the sofa?

Our Answer	Neural Image-QA	Ground Truth
Window	Window	Blinds

How many pillows are there on the sofa?

Our Answer	Neural Image-QA	Ground Truth
2	2	1

Fig. 11: Image Question pairs where both our model and Neural Image-QA model fails

What is the largest dark brown object in this picture?

Our Answer	Neural Image-QA	Ground Truth
bookshelf	table	table

What is the red object on the stove?

Our Answer	Neural Image-QA	Ground Truth
hot_water_heater	tea_kettle	tea_kettle

Fig. 12: Image Question pairs where Neural Image-QA model gives correct answer but our model predicts wrong answer. It can be observed that our model is mainly failing to answer those questions which that does not involve the questions involving depth

How many white boards are there?
Predicted Answer : 2

What is the Colour of Dustbin?
Predicted Answer : blue

How many doors are there?
Predicted Answer : 3

Fig. 13: Image question pairs where our model answers the questions on images taken at University of Michigan in Bob Betty Beyster hall

REFERENCES

[1] Dell Zhang, Wee Sun Lee. Web based pattern mining and matching approach to question answering In Proceedings of the 11^{th} Text REtrieval Conference

[2] Ravichandran D and Hovy E. Learning surface text patterns for a question answering system. *40^{th} Annual Meeting on Association of Computational Linguistics*, 2002, pp. 41-47

[3] J. Berant and P. Liang. Semantic parsing via paraphrasing.*ACL*, 2014.

[4] P. Liang, M. I. Jordan, and D. Klein. Learning dependency based compositional semantics. *Computational Linguistics*, 2013.

[5] M. Iyyer, J. Boyd-Graber, L. Claudino, R. Socher, and H. D. III. A neural network for factoid question answering over paragraphs. In EMNLP, 2014

[6] J. Weston, S. Chopra, and A. Bordes. Memory networks. *arXiv:1410.3916*, 2014

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012

[8] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *ECCV*, 2014

[9] M. Malinowski, M. Rohrbach and M. Fritz. Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. *ICCV*, 2015

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *arXiv:1409.0575*, 2014.

[11] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv:1409.4842*, 2014.

[13] Graves, A., Mohamed, A R. Hinton, G. Speech recognition with deep recurrent neural networks. *International Conference on Acoustics, Speech and Signal Processing*,2013.

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[15] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *In CVPR*, 2015.

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

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

[18] Ren, M.; Kiros, R.; and Zemel, R. S. 2015. Exploring models and data for image question answering. *arXiv 1505.02074*

[19] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. *In 3D Vision (3DV), 2016 Fourth International Conference on, pages 239248 IEEE*, 2016

[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, 2015.

[21] Nathan Silberman, Derek Hoiem, Pushmeet Kohli and Rob Fergus Indoor Segmentation and Support Inference from RGBD Images *ECCV* 2012

[22] Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. CNN Features off-the-shelf: an Astounding Baseline for Recognition. *CoRR, abs/1403.6382*, 2014.

[23] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. *In arXiv:1505.02074*. 2015.

[24] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia, Abc-cnn: An attention based convolutional neural network for visual question answering, *arXiv:1511.05960*, 2015.

[25] Wu, Z., and Palmer, M. 1994. Verb semantics and lexical selection. *In 32nd Annual Meeting of the Association for Computational Linguistics, 133138*