Visual Question Answering Through Adversarial Learning of Multi-modal Representation

Iqbal Chowdhury [1], Kien Nguyen Thanh [2], Clinton fookes [2], and Sridha Sridharan [2]

[1]Queensland University of Technology
[2]Affiliation not available

October 30, 2023

Abstract

Solving the Visual Question Answering (VQA) task is a step towards achieving human-like reasoning capability of the machines. This paper proposes an approach to learn multimodal feature representation with adversarial training. The purpose of the adversarial training allows the model to learn from standard fusion methods in an unsupervised manner. The discriminator model is equipped with a siamese combinatin of two standard fusion method namely multimodal compact bilinear pooling and multimodal tucker fusion. Output multimodal feature representation from generator is a resultant of graph convolutional operation. The resultant multimodal representation of the adversarial training allows the proposed model to infer the correct answers from open-ended natural language questions from the VQA 2.0 dataset. An overall accuracy of 69.86\% demonstrates the accuracy of the proposed model.

VISUAL QUESTION ANSWERING THROUGH ADVERSARIAL LEARNING OF MULTI-MODAL REPRESENTATION

Muhammad Iqbal Hasan Chowdhury, Kien Nguyen, Clinton Fookes, Sridha Sridharan

Queensland University of Technology
{m2.chowdhury,k.nguyenthanh, c.fookes, s.sridharan}@qut.edu.au

ABSTRACT

Solving the Visual Question Answering (VQA) task is a step towards achieving human-like reasoning capability of the machines. This paper proposes an approach to learn multimodal feature representation with adversarial training. The purpose of the adversarial training allows the model to learn from standard fusion methods in an unsupervised manner. The discriminator model is equipped with a siamese combinatin of two standard fusion method namely multimodal compact bilinear pooling and multimodal tucker fusion. Output multimodal feature representation from generator is a resultant of graph convolutional operation. The resultant multimodal representation of the adversarial training allows the proposed model to infer the correct answers from open-ended natural language questions from the VQA 2.0 dataset. An overall accuracy of 69.86% demonstrates the accuracy of the proposed model.

Index Terms— Visual Question Answering (VQA), Adversarial Learning, Multimodal Representation Learning, Scene Understanding.

1. INTRODUCTION

The ability to understand both the natural language sentences and visual data (image/video), and to further perform reasoning or decision making based on that is a good metric to measure machine's capability of human-like understanding of multimodal data. Machine learning models are trained on annotated image-question pairs to generate natural language answer for unseen image-question pairs. This paper attempts to learn a better multimodal representation with the use of Generative adversarial network[1] to bolster the performance of the VQA model.

Current approaches to solve the VQA task relies on first obtaining an encoding of the question sentence using recurrent architectures, e.g. recurrent neural network (RNN), long short-term memory[2] (LSTM) or gated recurrent unit (GRU). Also, covolutional neural network is used either to down-sample the feature size of the whole image or detected bounding boxes (by region proposal network, e.g. R-CNN[3]). Different variety of attention-mechanism and model structures of increasing complexity have been proposed to train the models on deep feature encodings of the textual and visual modality. A graph learning mechanism is used in[4], where semantic connection among detected bounding boxes are learned through graph convolution, which are conditioned on the question sentence encoding. Also, a relation-aware graph network is proposed[5] where different variety of relations among detected bounding box objects of the image are learned through a proposed relation encoder. However, we argue that the question sentence itself is the first and foremost component to define the context of reasoning for the VQA task. We believe that learning the context specific mulimodal representation is the key which drives the overall inference of the VQA model. All the existing approaches encode the question sentence as a whole and use that encoded representation in different model architectures. Existing approaches do not employ any mechanism to improve the obtained multimodal representation in an unsupervised manner.

GAN[1] models exhibit success in synthesizing new data from existing data distribution. The adversarial training approach is used in this paper to obtain an optimal multimodal feature representation where context specific multimodal features is generated and updated for the image and open-ended natural language question pair. The proposed approach in this paper allows the VQA model to further improve the multimodal feature representation through an adversarial training to bolster the performance of the VQA model. Figure 2 shows the overall architecture of the proposed model, where the image and question is forwarded through the generator to generate the fine grained multimodal feature representation throgh grapn convolutional LSTM. Later, the discriminator provides feedback whether the multimodal feature representation obtained through the generator is optimal or not. The discriminator network is a siamese combination of two standard fusion method, which allows the discriminator to provide more robust feedback on the performance of the generator network. The adversarial learning approach allows the model to update both the generator and the discriminator model over the time, which results in a better generator model to produce an optimal multimodal feature representation to predict the final answer word.

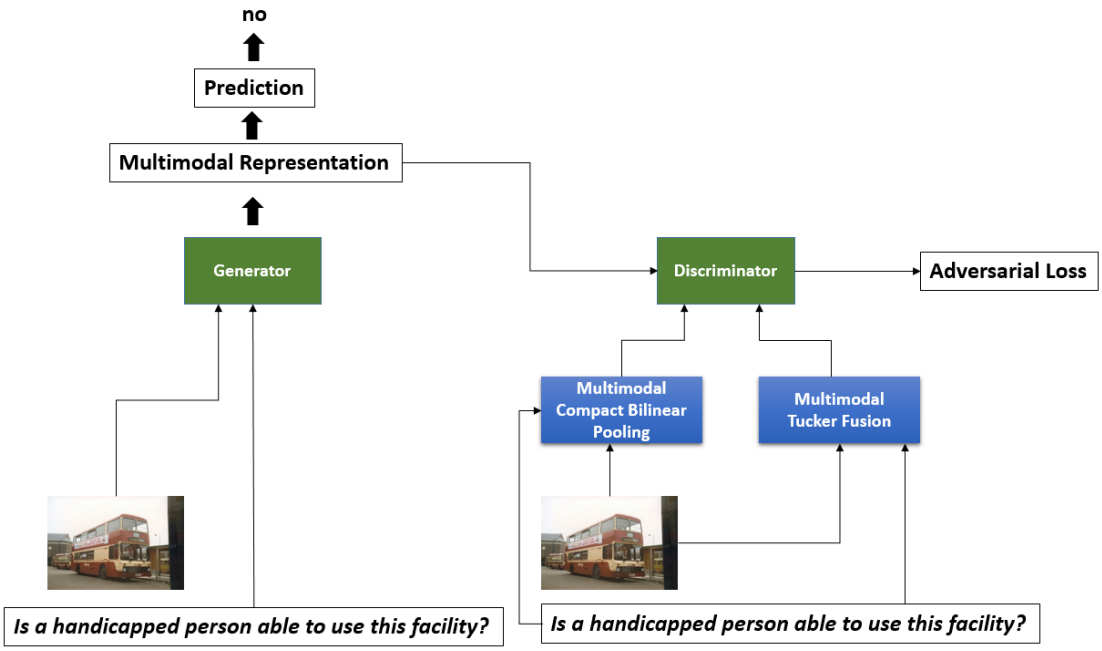

Fig. 1. An optimal representation of the multimodal feature is obtained with the use of GAN. Multmodal compact bilinear pooling and multimodal tucker fusion are used as the standard fusion method, where the generator loss tries to produce an optimal multimodal feature which is close to the multimodal feature representation of provided by the discriminator network. Proposed unsupervised learning of multimodal feature bolsters the performance of the VQA model.

The generator network of the proposed model uses graph convolution operation which is conditioned on the extracted phrase of the question sentence. Also, relationship among the detected object in the images are considered prior to the graph convolution operation. The output of the graph convolution conditioned on each of the phrase is passed through the LSTM, and finally the multimodal representation is found for the whole question sentence. The purpose of the discriminator network is examine the closeness of the derived multimodal feature representation with the original image and question features. A siamese combination of Multimodal compact bilinear pooling(MCBP)[6] and multimodal tucker fusion(MTF)[7] are used as the standard fusion method in the discriminator network. The adversarial loss penalize the generator in a mini-max game where the generator is forced to create better multimodal representation.

Our main contribution is the development of a VQA model which learns an optimal multimodal feature representation in an unsupervised manner. Competitive performance of the proposed method demonstrates the effectiveness of the proposed method. The proposed method uses object level feature to perform graph convolution operation, and the resultant intermediate feature representation is further improved with the help of adversarial minimax optimization.

2. RELATED WORK

Context specific multimodal representation learning is the main challenge for the VQA task. VQA problem was first studied by Mlinowski et al.[8] which combines semantic parsing and image segmentation. Geman et al.[9] Uses an automatic query generator which was trained on annotated images. Earlier approaches are limited on the form of questions that can be answered. Deep neural network architecture combined of CNN[10] and recurrent neural network (RNN) has become popular to learn the mapping from images to sentences. Usual approach is to use only image features with CNN but [11] uses multiple sources e.g. image content, generated image captions and mined external knowledge to feed to an RNN for VQA. Quality of information available in the KB is a vital fact for using knowledge bases with VQA. Hand crafted knowledge bases are too specific and on the other hand if constructed from Wikipedia or similar sources it becomes patchy and inconsistent. Wu. et. el.[11] proposes building a pre-build large-scale KB from which information can be extracted using a standard RDF query language.

Xiong et. el.[12] uses an input fusion layer to map the image patches with the dimensionality of the question vector. Both the attention and memory mechanism are incorporated in this deep neural network architecture. This approach is tested on three dataset namely, babI- 10k, DAQUAR-ALL vi-

sual dataset and Microsoft COCO dataset. Question answering is achieved with a multi-class classifier which is trained by using both the dense question embedding and image features. First questions are encoded with LSTMs and then question vectors are combined with image vectors by element wise multiplication. Most frequent answers are considered as possible outputs. This large scale open-ended VQA dataset is based on MS COCO [8]. CNN+BOW method is used as a baseline for this dataset which encodes image with CNN features and questions with BOW representation. Later a softmax neural network classifier is trained with a single hidden layer of 50 units and an output space being the 500 most frequent answers in the training set.

For image based question answering[13] proposes combination of internal representation of the content of image with information extracted from a general knowledge base. Deeper understanding of the scene is achieved by merging of textual information of the knowledge base and the textual representation of the semantic content of the images from Toronto COCO-QA[14] and MS COCO-VQA. More enriched knowledge bases will facilitate a robust visual question answering system.

Combination of LSTM for the query with a CNN for the image is proposed for VQA system in[15] for the DAQUAR dataset[16]. [15] prefers single word as the answer. A single recurrent network is used to perform both encoding and decoding. Attention based encoder-decoder model is used for VQA in[17]. [17] uses one or multiple image regions to determine answer for the visual question answering models. They demonstrated how spatial mechanism works to read a picture over the DAQUAR[18] and VQA[19] dataset. Question is used to compute an attention over the input image. Later answer is computed based on the question and the attended image features.

Syntactic supervision is found in the form of dependency trees for question answering in[20]. Models in[20] are inspired by neural modules and evaluates knowledgebase reasoning and visual question answering. Vision models in[20] uses reinforcement learning technique to backpropagate through a sampling mechanism for the visual question answering 19task. [20] produces attention maps to answer object reference questions through parsing the question sentence.

Also, recent approaches incorporates graph convolution[5][4] operation with the attention mechanism in a aim to capture the interactions among objects in the visual scene. Also, an attempt to find relation among among objects through pairwise relational reasoning is explored in[21]. However, none of the existing approaches focus on dividing the question sentence intor meaningful chunks i.e. phrases to better understand the context of the reasoning. This paper advocates the fact that the question sentence is the main agent of driving the context of reasoning for the VQA task. The proposed method in this paper divides the question sentence into phrase-based conceptual unit to further perform reasoning of the VQA task, whic reflects the usefulness of the divide-and-conquer style approach to solve the VQA task. Moreover, the phrase-conditioned graph convolution is further strengthened by the incorporation of relational features with the convolutional features of the nodes of the grpah convolution operation.

3. MULTIMODAL FEATURE GENERATION BY GRAPH CONVOLUTIONAL LSTM

Many types of feature fusion techniques have been proposed, e.g. multimodal compact bilinear pooling[6], multimodal tucker fusion[7] etc. These feature fusion methods demonstrate promising performance in the supervised learning setup for the ImageQA models. However, how to learn from these standard fusion methods in an unsupervised manner for the VQA task is still a less explored area. An unsupervised approach to learning from multiple standard fusion techniques is proposed in this chapter. A generative adversarial network-based training is proposed where the generator network generates question sentence phrase conditioned features through graph convolutional LSTM. Also, the discriminator network is a Siamese[22] network of standard multimodal fusion techniques, which compares the generator network's output with the siamese combination of standard fusion methods to minimize an adversarial loss.

An encoding of the question sentence is obtained by using recurrent architectures, e.g. recurrent neural network (RNN), long short-term memory (LSTM)[2] or gru. In the traditional approach obtaining the question, sentence embedding is considered as the first step to solve the VQA task. Also, the convolutional neural network is used either to down-sample the feature size of the whole image or detected bounding boxes (by region proposal network, e.g. R-CNN[3]). Different variety of attention-mechanism and model structures of increasing complexity have been proposed to train the models on deep feature encoding of the textual and visual modality. A graph learning mechanism is proposed in[4]. In the graph learning mechanism, the semantic connection among detected bounding boxes is learned through graph convolution, which is conditioned on the question sentence encoding. Also, a relation-aware graph network is proposed[5], where a variety of different relations among detected bounding box objects of the image are learned through a proposed relation encoder. However, we argue that the question sentence itself is the first and foremost component to define the context of reasoning for the VQA task. We believe that learning the context-specific multimodal representation is the key that drives the overall inference of the VQA model. All the existing approaches encode the question sentence as a whole and use that encoded representation in different model architectures. Existing approaches do not employ any mechanism to improve the obtained multimodal representation in an unsupervised manner.

GAN[1] models exhibit success in synthesizing new data

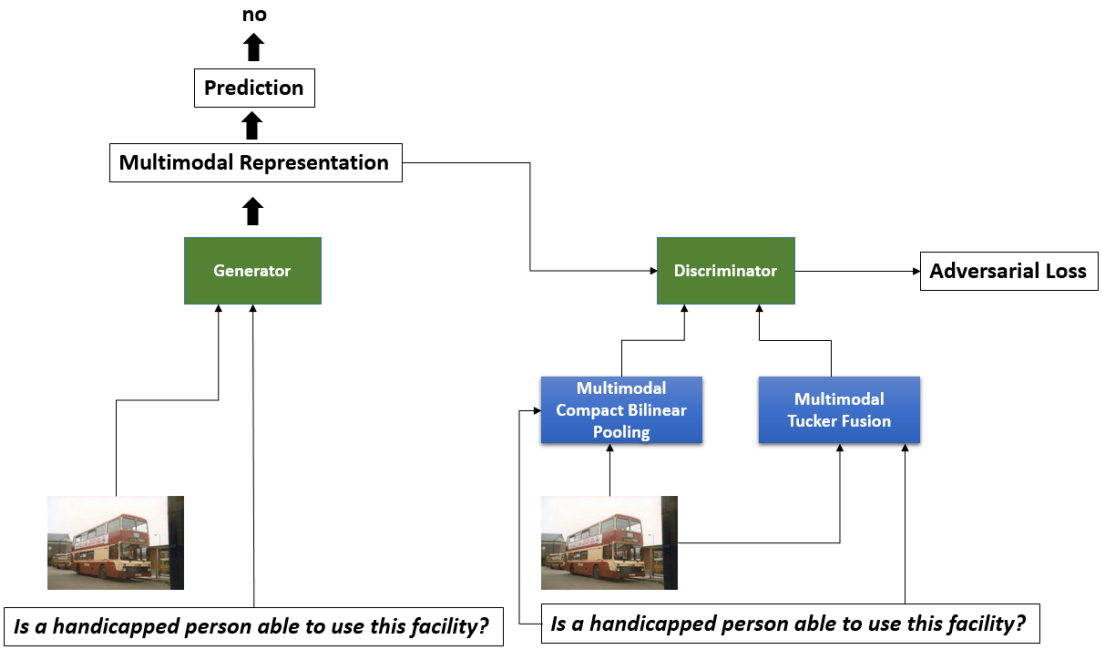

Fig. 2. An optimal representation of the multimodal feature is obtained with the use of adversarial training. Multimodal compact bilinear pooling and multimodal tucker fusion are used as the standard fusion method, where the generator loss tries to produce an optimal multimodal feature, which is close to the multimodal feature representation provided by the discriminator network. Proposed unsupervised learning of multimodal feature bolsters the performance of the VQA model.

from the existing data distribution. The adversarial training approach is used in this chapter to obtain an optimal multimodal feature representation where context-specific multimodal features are generated and updated for the image and open-ended natural language question pair. The proposed approach in this chapter allows the VQA model to further improve the multimodal feature representation through adversarial training to bolster the performance of the VQA model. Figure 2 shows the overall architecture of the proposed model, where the image and question are forwarded through the generator to generate the fine-grained multimodal feature representation through graph convolutional LSTM. Later, the discriminator provides feedback on whether the multimodal feature representation obtained through the generator is optimal or not. The discriminator network is a Siamese combination of two standard fusion methods, which allows the discriminator to provide more robust feedback on the performance of the generator network. The adversarial learning approach allows the model to update both the generator and the discriminator model over time, which results in a better generator model to produce an optimal multimodal feature representation to predict the final answer word.

The generator network of the proposed model uses graph convolution operation, which is conditioned on the extracted phrase of the question sentence. Also, the relationship between the detected object in the images is considered before the graph convolution operation. The output of the graph convolution conditioned on each of the phrases is passed through the LSTM, and finally, the multimodal representation is found for the whole question sentence. The purpose of the discriminator network is to examine the closeness of the derived multimodal feature representation with the original image and question features. A Siamese combination of mcbp[6] and mtf[7] is used as the standard fusion method in the discriminator network. The adversarial loss penalizes the generator in a mini-max game where the generator is forced to create better multimodal representation.

The proposed model is equipped with a generator and discriminator network, which are glued together to jointly optimize the adversarial loss and answer prediction loss. The generator network uses graph convolutional LSTM to generate context-specific multimodal feature representation. The discriminator network compares the generator output with the output of a Siamese combination of two standard fusion methods. The generator network is described in this section. The discriminator network for adversarial training is described in the next section.

The generator serves the purpose of generating context-specific multimodal feature representation. The open-ended natural language question is the main catalyst to define and influence the overall reasoning process of the VQA task. Thus, the generator network considers the question sentence

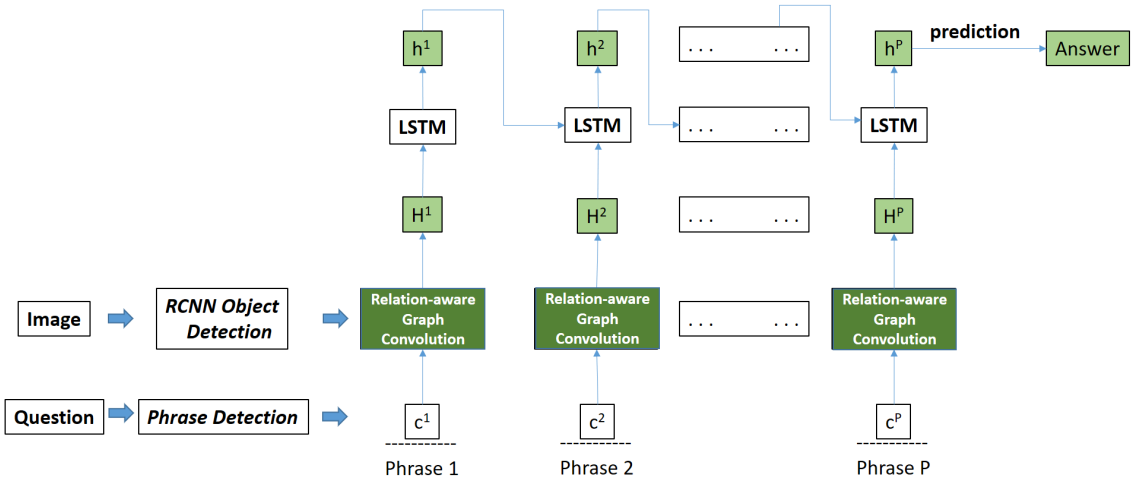

Fig. 3. In the generator network, each of the question phrases is considered individually instead of the whole question sentence encoding. Based on each phrase the graph convolution operation is performed on image features. Detected bounding boxes of images are considered as nodes in the graph convolution. Convoluted graph convolution features conditioned on each of the phrases are then passed through LSTM. The final answer prediction is performed when all the phrases are addressed for the graph convolution operation.

in a phrase-by-phrase manner to derive context-specific visual feature representation through graph convolution operation. The graph convolution operation also considers the object-level relations and their feature representation to obtain a better multimodal feature representation. The following subsections describe the phrase parsing and phrase conditioned graph convolution operation, which results in the multimodal feature representation output of the generator network.

The first contribution of the proposed method is to divide the natural language sentence into meaningful chunks to bolster the human-like reasoning capability of the model through graph convolutional LSTM. If any natural language sentence is distilled down, then phrases are the most meaningful conceptual unit available. These chunks give a hint of step-by-step evolution of the understanding for the whole question sentence. Thus, the VQA models must understand and correlate the understanding of the visual data (images) conditioned on the phrases to achieve better reasoning capability of the model. As visible in Figure 3, each of the phrases chunks c^1 to c^P is used to condition the relation-aware graph convolution operation, where nodes of the graph convolution are equipped with relational features associated with every detected bounding box of the image. Graph convolution operation conditioned on each of the frames results in the outputs H^1 to H^P. The LSTM memory is updated with the output of graph convolution operations started from the conditioning of the very first phrase c^1 till the encounter of the last phrase c^P.

$$Q = [c^1, c^2, ..., c^p], \qquad p = 1, 2, ..., P \tag{1}$$

Standard NLTK[23] parser is used during implementation to extract meaningful chunks i.e. the phrases of the question sentence. Equation 1 lists the different chunks, which are being extracted from the question sentence Q. There is P number of phrases based on the length of the open-ended questions. Each subscript of the element c refers to the sequence number of phrases extracted from the question sentence Q. According to the sequence of extraction, each of the phrase chunks is used to condition the subsequent steps of relation-aware graph convolution operations and LSTM state updates.

3.1. Relation aware and Phrase Conditioned Graph Convolution

The second contribution of the proposed method uses relation aware visual feature representation of each bounding box as nodes in the subsequent graph convolution operation. On the contrary, the proposed graph convolution operation described in[4] directly uses the average of CNN features of each bounding box as the node to build the adjacency matrix for the subsequent graph convolution operation. Before the construction of the adjacency matrix, a relation aware feature representation for each of the bounding boxes is obtained by following the principle to capture relations among objects as proposed in[24]. Pairwise relation is computed for each of the detected bounding boxes as shown in Equation 12. The final feature representation v_n for each of the bounding is obtained by concatenating the relation features v_n^* with the original representation of v_n as shown in Equation 3. Then

again, extracted phrase c^p is concatenated with v_n as shown in Equation 4, which is then used to build the adjacency matrix for the subsequent graph convolution operation. To elaborate further, let us consider feature representation v_1 for bounding box 1 among the N number of detected bounding boxes of the respective image. The next step is to create a set of all possible pairs with other detected bounding boxes of the image, which results in a sequence starting from (v_1, v_2, c^p) then (v_1, v_3, c^p) , and then up to (v_1, v_N, c^p). It is to be noted that the relation aware graph convolution operation is performed for each of the extracted phrases of the question sentence, and the output of each step is propagated through LSTM towards the final answer word prediction.

$$\begin{aligned} v_n^* = \sum_{i,j} g_\theta(v_n, v_z, c^p) \\ where, n = 1, 2, ..., N \\ z = 1, 2, ..., N \\ and\ p = 1, 2, ..., P \end{aligned} \tag{2}$$

$$v_n = v_n || v_n^* \tag{3}$$

The graph convolution operation follows a similar principle as described in[4]. The main contribution of the model described in[4] is to learn an adjacency matrix for an undirected graph where edges of the respective graph are conditioned on the question sentence as a whole. Each of the detected bounding boxes is considered as a vertex of the graph structure. Then the average of the convolutional feature of each bounding is concatenated with the question embedding. Unlike the approach of using the encoding of the whole question sentence, the proposed methodology of this chapter uses each conceptual unit i.e. the phrases separately to build multiple adjacency matrices. Also, nodes of the graph are made relation-aware with the inclusion of relational features.

We aim to build an undirected graph structure for each conceptual unit i.e. the phrase chunks and to further extract feature representation through graph convolution for each of the phrase chunks. Thus, for each of the phrases, there will be an adjacency matrix building operation for each of the undirected graphs, G^p, where $p = 1, 2, ..., P$. To learn the adjacency matrix of each phrase i.e. to build an undirected graph $G^p = \{V^p, E^p, A^p\}$, the respective averaging of convolution operation of each bounding box area and the concatenation with word embedded encoding of each phrase results in P number of undirected graphs and respective graph edges, where P is the number of extracted phrases from the question sentence Q.

Again, for each of the undirected graphs, their adjacency matrix $A^p \in R^{N \times N}$, where N denotes the number of detected bounding boxes, i.e. $|v| = N$. For each graph of the G^p, the number of detected bounding boxes is considered as the number of vertices to build the respective adjacency matrix A^p. In each of the adjacency matrix A^p, the respective relation-aware visual features, v_n (obtained by the Equation 3) of bounding boxes, are concatenated with the respective chunk embedding, c^p. The operation in Equation 4 allows each adjacency matrix to hold and map the similarities between feature vectors along with the relevance of the visual features to specific conceptual units i.e. the phrase chunks. Each edge $(i, j, A^p_{i,j} \in E^p)$ of the graph G^p is conditioned on the phrase chunks c^p as obtained by the conditioning operation in Equation 4. The conditioning operation in Equation 4 is similar to the method proposed in[4]. However, the difference we are proposing is to follow a divide-and-conquer-based approach, where we divided the question sentence into meaningful chunks to further achieve the capability of human-like reasoning. Also, the feature representation of each of the nodes in the undirected graph contains relational features information. The graph convolution output for each of the phrase chunks is further passed through an LSTM, which is described in Section 3.2 to consolidate the output of each phrase chunk based graph convolution to the final answer word. In Equation 4, F^p: $R^{d_v^p + d_q^p} \rightarrow R^{d_e^p}$ is a non-linear operation, which makes the concatenation operation suitable for differentiation in the neural network operation. Also, F^p consists of two 2 dense layers of size 512. Again, d_v^p, d_q^p and d_e^p are the dimensions of the image, question and the joint embedding vector for each of the respective phrase chunk $p = 1, 2, ..., P$. The matrix $E^p \in R^{N \times d_e^p}$ holds the embedding e_n^p, which results in the formation of adjacency matrix $A^p = E^p(E^p)^T$ respective to each of the phrase chunk p. Similar to the approach of[4], the question specific graph structure is obtained by adopting the ranking strategy $N(i) = topm(a_i^p)$, where m largest values of the ith row of the adjacency matrix A^p is returned. To perform graph convolution operation a patch operator is defined as $f_k(i) = \sum_{j \in N(i)} w_k(u(i,j)) v_j \alpha_{ij}$, where, $k = 1, 2, ..., K$. The terms, w_k and α_{ij} refer to the set of learnable weights, which are finally put together in the matrix $G_k \in R^{\frac{d_h}{k} \times d_v}$, where h is the chosen dimensions of the output of graph convolution. Equation 5 refers to the concatenation operation over K kernels, which results in the output h_i of the graph convolution operation at vertex i. A max-pooling operation across convolved features $H \in R^{N \times d_h}$ results in the output, H^p (of the phrase conditioned graph), which is further passed through LSTM.

$$\begin{aligned} e_n^p = F^p([v_n || c^p]) \quad p = 1, 2, ..., P \\ and, n = 1, 2, ..., N \end{aligned} \tag{4}$$

$$h_i = \|_{k=1}^{K} G_k f_k(i) \tag{5}$$

3.2. Graph Convolutional LSTM

Phrase-conditioned and relation-aware graph convolution give the proposed model a better ability to understand the

context in a divide and conquer manner. A graph recurrent network[25] has been used for other tasks such as traffic data prediction. In the proposed method of this chapter, the graph convolutional LSTM is uniquely applied for the VQA task. However, it is essential to attend to all different conceptual units, and also to maintain the relevancy gradually over time. An LSTM cell is used to maintain and hold the contextual relevance and connection of the graph convolution operation for each of the phrase chunks. In this manner, the output of graph convolution conditioned on the first phrase chunk is passed through the LSTM to the next detected phrase chunk conditioned graph convolution operation. The process of the LSTM state update is continued until all the phrases are considered for the graph convolution operation. The output of the very last LSTM hidden state representation is used as the vector to predict the final answer word.

$$\begin{aligned} f_t &= \sigma(w_f H_t^p + u_f h_{t-1}^p + b_f) \\ i_t &= \sigma(w_i H_t^p) + u_i h_{t-1}^p + b_i) \\ o_t &= \sigma(w_o H_t^p + u_o h_{t-1}^p + b_o) \\ \tilde{c_t} &= tanh(w_c H_t^p + u_c h_{t-1}^p + b_c) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c_t} \\ h_t &= o_t \odot tanh(c_t) \end{aligned} \tag{6}$$

Ther terms f_t, i_t and o_t respectively refers to the forget gate, input gate and output gate of the LSTM operation. The output of the graph convolution H_t^p is used as the input to the LSTM cell, where, the superscript $p = 1, 2, ..., P$ refers to the respective phrase of the question sentence. Again, w, u and b terms refer to respective weight and bias values associated with the graph convolution input and the hidden state of the LSTM. Sigmoid activation is used to maintain the non-linearity for the operation in the forget, input and output gates. Also, hyperbolic tangent activation is used for the cell state update. Figure 3 shows the hidden state (h^1 to h^P) propagation of the graph convolution operation through the LSTM state update, which are conditioned respectively on the phrases c^1 to c^P.

4. MULTIMODAL FEATURE DISCRIMINATOR FOR VQA

The discriminator network consists of two standard fusion methods, namely MCBP and MTF. Figure 4 shows the Siamese combination of these two standard fusion methods. Image and question embedding are passed through each of the fusion methods. The outputs of each of the fusion methods are passed through LSTMs, which are then combined based on similarity. The combined representation is again passed through an LSTM. The resultant LSTM output of the discriminator network is used to measure the multimodal feature representation of the generator network.

$$\begin{aligned} M_1 &= MCBP(I, Q) \\ M_2 &= MTF(I, Q) \end{aligned} \tag{7}$$

$$\begin{aligned} s &= exp(-||M_1, M_2||) \\ d &= exp(-||s, h^p||) \end{aligned} \tag{8}$$

Discriminator network compares the multimodal feature representation produced by the generator network with the feature representation provided by the MCBP operation. The MCBP is used as the standard method for feature fusion and considered as the ground truth feature representation in the proposed method. In Equation 7, the M_1 and M_2 respectively refer to the output of the MCBP and the MTF fusion methods. The output of the MCBP and MTF fusion methods are compared in Equation 8, which results in the output of s. Later, s is again compared with the output of the generator network, which is then classified as either relevant or not with Siamese representation of feature fusion.

5. ANSWER PREDICTION THROUGH ADVERSARIAL TRAINING

The output of the LSTM hidden state is considered as a final resultant vector to be compared for a multi-class classification setup. In the experimental setting, each of the most frequent answers is considered as an individual class for the VQA problem. Binary cross-entropy loss is used in the experiment, which is similar to[4], and the function is given in Equation 7. The term y in the equation refers to the output of the LSTM hidden state, i.e. $y = h_t$. Also, t is the soft target score for each of the answer class. Equation 9 and 10 represent the loss function of the adversarial training of the proposed model.

$$L(G, D) = L_{GAN}(G, D) + L_{QA}(G) \tag{9}$$

$$\begin{aligned} L_{GAN} = \min_G \max_D & log(dLSTM(M_1, M_2)) \\ & + log(1 - gLSTM(h)) \end{aligned} \tag{10}$$

$$\begin{aligned} L(t, y) &= \sum_i t_i \log(1/(1 + \exp(-y_i)) \\ &+ (1 - t_i) \log(\exp(-y_i)/(1 + exp(-y_i)) \end{aligned} \tag{11}$$

6. EXPERIMENTAL RESULTS

Classification accuracy is considered as the evaluation metric during the experimentation. Also, inter-human variability is considered as shown in Equation 8.

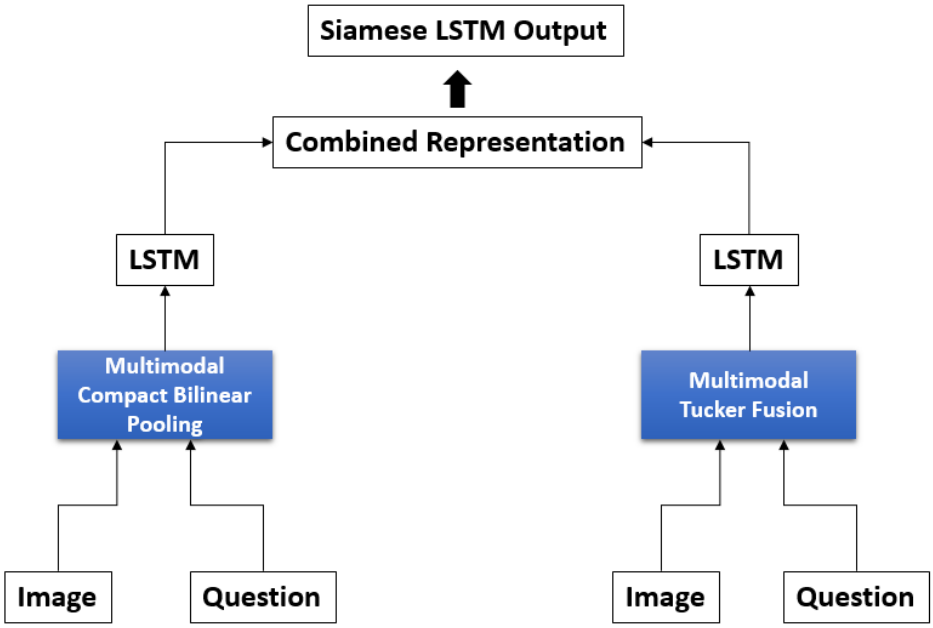

Fig. 4. A Siamese combination of two standard fusion methods in the discriminator network. Image and questions are respectively considered with the MCBP and MTF fusion method. The output of fusion methods is passed through LSTM and the similarity between these fused representations are measured, which results in the final combined representation, which is then used to compare the closeness of the generated multimodal feature representation.

Fig. 5. Qualitative output of the proposed method. Final answer prediction is performed based on the optimal multimodal representation which is obtained through adversarial training. The activated region in the images refer to some node activation in the graph convolution operation of the generator network.

Methods	Y/N	Number	Other	Overall
ReasonNet[26]	73.86	41.98	57.39	64.61
Bottom-Up[27]	82.20	43.90	56.26	65.67
Counting Module[28]	83.56	51.39	59.11	68.41
Graph Conv.[4]	82.91	47.13	56.22	66.18
Murel[21]				68.41
Proposed Method	**85.07**	**56.14**	**59.32**	**69.86**

Table 1. Comparison of accuracy on different answer types of the VQA 2.0 dataset. The bold row in the table shows the performance of the proposed method.

6.1. Evaluatoin Metric

The open-ended question may have different right answers. Thus, considering a single answer as the fixed class may lead the classifier to consider a variant of the right answer as wrong. To overcome this problem, the authors in[29] propose to consider 10 answers for every single question in the VQA2.0 dataset. A predicted answer is only considered correct if at least 3 of the 10 available answer matche with the predicted answer as shown in Equation 12. This metric allows the trained models to overcome the limitation of classification based approaches.

6.2. Experimental settings

The experiment is carried on the VQA 2.0 dataset. This dataset contains open-ended questions associated with 10 answers annotated by human users. A total of 1,105,904 questions are paired with 204,721 images from the Microsoft COCO dataset. Then, 40% of the questions are used for training, with 40% and 20% used respectively for the testing and validation.

In the generator network, bottom-up features are used where each image is represented as a set of 36 localized regions resulting in convolutional feature vectors of dimension 2048. Again, following the strategy of[4] the bounding box corners are normalized to remain in the interval of $[0, 1]$ and the features vector size results in the dimension of 2052. Again, the textual question sentence is encoded with a dynamic GRU with a hidden state size of 1024. The pre-trained GLoVe embedding is used to extract a 300 dimensional embedded representation of each of the question sentence tokens. In all the dense and convolutional layers the ReLU is used to ensure the non-linearity of the model. The loss function is optimized for 35 epochs with Adam optimizer with a learning rate of 0.0001. Also, 2 spatial graph convolutional layers of dimension 2048 and 1024 are used for the graph convolution. Following a similar approach to[4], the neighbourhood size and the number of kernels are respectively chosen as 16 and 8 for the experimentation. Again, two-layer LSTM is used in the discriminator network where each layer consists of 2048 hidden units.

Methods	Overall Accuracy
MCBP only	64.61
MTF only	65.67

Table 2. Accuracy of the two ablation instances on the validation split of the VQA 2.0 dataset.

6.3. Experimental Output

Accuracy (as shown in Equation 8) is used as the evaluation metric for the proposed model. Accuracy is compared within the 'Test-standard' split of the dataset, and the proposed model is trained on the 'TrainVal' split of the VQA 2.0 dataset. Table 1 shows the accuracy of different answer types of the VQA 2.0 dataset. It is evident that the proposed model exhibits competitive performance and is pointing in a new direction to improve the interpret-ability of the VQA models, by considering the question sentence in a divide and conquer manner with the relational features equipped graph convolutional LSTM operation. Two ablation instances are considered as part of the ablation study for the proposed model. One instance considers only the MCBP fusion in the discriminator network, and the other instance considers only the MTF fusion method in the discriminator network. Table 2 shows the performance for these two ablation instances.

$$Accuracy(answer) = \min(1, \frac{no.\,of annotated\,answer}{3}) \tag{12}$$

Figure 5 shows the qualitative output of the proposed model. As seen, original images are passed through the proposed relation-aware and phrase conditioned graph convolution, and the output images exhibit the most dominant nodes (highlighted) of the relation-aware and phrase-conditioned graph convolution operation. It is visible that the network focuses on the most important regions, which influence the overall feature representation based on the context of the test question sentence.

7. CONCLUSION AND FUTURE WORKS

The proposed phrase-conditioned and relation-aware graph convolutional LSTM method considers the main question sentence for the VQA task in a more interpretable manner with the extraction of phrase/coceptual chunk. Also, nodes of the graph convolution operation includes relation information among objects, which results in better formation of the adjacency matrix for the subsequent graph convolution operation. Future endeavour will include multimodal attention mechanism to further improve the accuracy of the proposed approach.

8. REFERENCES

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[2] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins, "Learning to forget: Continual prediction with lstm," 1999.

[3] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[4] Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot, "Learning conditioned graph structures for interpretable visual question answering," in *Advances in Neural Information Processing Systems*, 2018, pp. 8334–8343.

[5] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu, "Relation-aware graph attention network for visual question answering," *arXiv preprint arXiv:1903.12314*, 2019.

[6] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[7] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome, "Mutan: Multimodal tucker fusion for visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2612–2620.

[8] Mateusz Malinowski and Mario Fritz, "Towards a visual turing challenge," *CoRR*, vol. abs/1410.8027, 2014.

[9] Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes, "Visual turing test for computer vision systems," *Proceedings of the National Academy of Sciences*, vol. 112, no. 12, pp. 3618–3623, 2015.

[10] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, 2014, pp. 647–655.

[11] Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, and Anthony R. Dick, "Ask me anything: Free-form visual question answering based on knowledge from external sources.," *CoRR*, vol. abs/1511.06973, 2015.

[12] Caiming Xiong, Stephen Merity, and Richard Socher, "Dynamic memory networks for visual and textual question answering," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, 2016, pp. 2397–2406.

[13] Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, and Anthony R. Dick, "Ask me anything: Free-form visual question answering based on knowledge from external sources," *CoRR*, vol. abs/1511.06973, 2015.

[14] Fereshteh Sadeghi, Santosh K Divvala, and Ali Farhadi, "Viske: Visual knowledge extraction and question answering by visual verification of relation phrases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1456–1464.

[15] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015, pp. 1–9.

[16] Mateusz Malinowski and Mario Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2014, pp. 1682–1690.

[17] Huijuan Xu and Kate Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," *CoRR*, vol. abs/1511.05234, 2015.

[18] Mateusz Malinowski and Mario Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *Advances in Neural Information Processing Systems 27: Annual Conference on*

Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, 2014, pp. 1682–1690.

[19] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh, "VQA: visual question answering," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015, pp. 2425–2433.

[20] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein, "Learning to compose neural networks for question answering," in *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, 2016, pp. 1545–1554.

[21] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome, "Murel: Multimodal relational reasoning for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1989–1998.

[22] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang, "A siamese long short-term memory architecture for human re-identification," in *European conference on computer vision*. Springer, 2016, pp. 135–153.

[23] Edward Loper and Steven Bird, "Nltk: the natural language toolkit," *arXiv preprint cs/0205028*, 2002.

[24] Adam Santoro, David Raposo, David GT Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap, "A simple neural network module for relational reasoning," *arXiv preprint arXiv:1706.01427*, 2017.

[25] Zhiyong Cui, Kristian Henrickson, Ruimin Ke, and Yinhai Wang, "Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting," *arXiv preprint arXiv:1802.07007*, 2018.

[26] Ilija Ilievski and Jiashi Feng, "Multimodal learning and reasoning for visual question answering," in *Advances in Neural Information Processing Systems*, 2017, pp. 551–562.

[27] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4223–4232.

[28] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett, "Learning to count objects in natural images for visual question answering," *arXiv preprint arXiv:1802.05766*, 2018.

[29] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.