Inteligencia Artificial 27(73) (2024), 111-128
doi: 10.4114/intartif.vol27iss73pp111-128

INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

TRANS-VQA: Fully Transformer-Based Image Question-Answering Model Using Question-guided Vision Attention

Dipali Koshti[1], Ashutosh Gupta[2], Mukesh Kalla[3], Arvind Sharma[4]

[1,2,3] Department of Computer Science and Engineering, Sir Padampat Singhania University, Udaipur, Rajasthan, India.
[4] Department of Mathematics, Sir Padampat Singhania University, Udaipur, Rajasthan, India.

[1]dipali.koshti@spsu.ac.in, [2]ashu.gupta@spsu.ac.in, [3]mukesh.kalla@spsu.ac.in, [4]sharma.arvind@spsu.ac.in

Abstract Understanding multiple modalities and relating them is an easy task for humans. But for machines, this is a stimulating task. One such multi-modal reasoning task is Visual question answering which demands the machine to produce an answer for the natural language query asked based on the given image. Although plenty of work is done in this field, there is still a challenge of improving the answer prediction ability of the model and breaching human accuracy. A novel model for answering image-based questions based on a transformer has been proposed. The proposed model is a fully Transformer-based architecture that utilizes the power of a transformer for extracting language features as well as for performing joint understanding of question and image features. The proposed VQA model utilizes F-RCNN for image feature extraction. The retrieved language features and object-level image features are fed to a decoder inspired by the Bi-Directional Encoder Representation Transformer - BERT architecture that learns jointly the image characteristics directed by the question characteristics and rich representations of the image features are obtained. Extensive experimentation has been carried out to observe the effect of various hyperparameters on the performance of the model. The experimental results demonstrate that the model's ability to predict the answer increases with the increase in the number of layers in the transformer's encoder and decoder. The proposed model improves upon the previous models and is highly scalable due to the introduction of the BERT. Our best model reports **72.31%** accuracy on the test-standard split of the VQAv2 dataset.

Keywords: Question-guided VQA, Visual question answering, Transformer-based VQA, BERT-based VQA.

1 Introduction

Humans can easily understand the image and perform reasoning over the image content. Given an image, humans can easily perform complex tasks such as scene detection, object identification, Object counting, etc. However, performing such reasoning over the image is challenging for a machine. Visual question answering is nothing but developing an intelligent machine model that is capable of generating the correct answer for any query posed for the given image. Since the query is about the image, both the question and the image need to be processed in correlation. Such a problem, in the context of machine learning, is called multi-modal since it involves processing two different modalities: image and language.

ISSN: 1137-3601 (print), 1988-3064 (on-line)

Although there has been a lot of study on VQA problems, it remains difficult to increase the model's accuracy and surpass human accuracy. Any VQA model typically consists of four phases. The extraction of picture characteristics comes first, followed by the extraction of question features. The third step must do a joint representation of these two elements because the inquiry relates to the image. In the final step, the prediction of the answer is done. Early literature focused on extracting better image features by using various CNN models like GoogleNet[1-3], VGGNet [4-6], ResNet [7-9] etc. or extracting better question features using various word embedding techniques like word2Vec[10-11], Skip thought[8,12], LSTM [13-16] and used simple fusion methods such as element-wise addition or element-wise multiplication [7],[10],[17]. Recent literature explores improving the joint learning of image and question characteristics by using various attention models. Attention-based VQA models attend only to the dominant parts of an image or a question thereby improving the performance. Also, much recent literatures explored the transformer-based VQA models [18-20]. The transformers process the entire sentence at once as opposed to conventional word embedding algorithms, which break down the sentence into its parts and learn the context. The contribution of this research is listed here.

- An improved visual question-answering model based on a transformer (BERT) has been proposed.
- An extensive ablation study on the selection of hyperparameters of the model has been performed.
- The proposed model is a fully Transformer-based architecture that utilizes a transformer for question feature extraction as well as for extraction of question–guided image features.
- The suggested model uses an encoder-decoder strategy of BERT for deeper co-attention learning between question and image characteristics.

Due to the utilization of BERT, experimental findings show that the suggested model outperforms earlier SoTA models and is extremely scalable. Our top single model reports **72.31%** overall accuracy on the test-standard split of the VQAv2 dataset. We execute thorough ablation to examine the effects of different hyperparameter settings on the model's performance.

The rest of the sections are arranged as mentioned: The prior research in the area has been covered in Section 2, and the proposed TRANS-VQA framework and the working of its individual modules are explained in Section 3. The experimental design and findings are covered in Section 4 and finally, Section 5 summarizes and concludes the research work.

2 Prior Work

Vision–language tasks have nowadays attracted many researchers due to their challenges in processing and understanding multi-modal data. Visual question answering is such a type of vision–language task that demands deep comprehension of vision and language together. The VQA model is expected to explore dense interaction between the vision and language features efficiently to increase the answer prediction ability.

In [21] Yang et al. presented an attention-based method called stacked Attention Network. The authors proved that performing reasoning multiple times on images gives better results. They introduced an attention model where they ask the image a query multiple times to generate the answer. Lianli Gao et al. [22] suggested that the question-guided image features are richer and provide better content with respect to question features. The model extracts semantics from the question and important objects from an image (using an object detection algorithm). The model selects only question-related object regions and finally optimizes the QLOB attention and the language model by using a Softmax classifier to generate the correct response. In [23] authors presented a dense co-attention network for fusing question and image features. Each word in a question attends to every region in an image, and every region in an image attend to every word in a question, according to a fully bidirectional attention mechanism that was presented. Further stacking was used to perform multi-step reasoning. Rahman et al. [24] presented an improved attention mechanism called MCAoAN that captures intra and inter-channel attention. Also, they proposed two improved fusion techniques: multi-channel attention fusion and Multi-channel Mutan fusion. In [25] Nguyen et al. extracted the predicates from both the question and the image. These predicate features along with the image and question vectors are used for course-grained learning and the filtered features are used for fine-grained learning. Filtering removes unnecessary information from image and question features. ViLBERT [26] extended the BERT to jointly represent the text and image. The architecture has two streams: vision and language stream. In the visual stream, the image-guided language attention is performed and in the language stream, the language-guided image attention is done. [27] proposed a self–adaptive transformer-based VQA model that addresses the dynamic nature of question comprehension and dynamically discards the module that performs badly by considering the intermediate results of each module during the reasoning. Yan Zhang [28] addressed the object counting problem in VQA. Almost all the VQA models struggle with the Number questions in VQA. The model discards the redundant counting by removing overlapping object proposals to increase counting accuracy.

Authors of the MRA-NET model [29] explore two additional relational attention modules: Binary relation and ternary relation. In [30] Guo W. et al. claimed that attending to the answer is equally important as attending to the question and image to predict the correct answer. The answer information was used to correct the generated visual attention map. [31] F. Liu, J, proved that not only the question-guided image attention is important but also dense interaction between the words of the question is required to increase model performance. The model performs two layers of self-attention on the question and this attended question is then used for guiding image attention.

3 Methodology

The VQA task can be formally defined as follows. Let ‘I’ be the Image, ‘S’ be the corresponding question sentence, and the answer ‘a’ is to be predicted.

$$a = \arg\max_{ans \in A} P(ans|(I,S); \theta) \tag{1}$$

In eq (1), ‘P’ is the probability of an answer ‘ans’ from the answer set A, given a pair of images ‘I’ and question ‘S’. Here, A is an answer set and θ denotes the learnable parameters. We would like to maximize this probability. This section discusses the proposed model and each component. Figure 2 demonstrates the detailed design of the TRANS-VQA model. We used the VQAv2 dataset for training and testing our model.

3.1 The Dataset

VQA2.0 dataset [17] is created through human labelling and uses images from the COCO image caption dataset. The VQA dataset contains three splits - training set, validation set as well as test set. Each set contains the image, related questions, and the answers. Overall 204K images were taken from the COCO dataset. Corresponding to each image there are 3 questions. In turn, there are 10 human-labelled answers associated with each question. Thus, there are approximately 614K free-form questions and 1 billion answers. This is the largest dataset available to date for solving the VQA problem.

3.2 Pre-processing

Before we feed information into the model, a great deal of pre-processing work is needed. During the training phase, the pre-processed questions, images, and answers are fed into the model. The questions, the answers, and the images were pre-processed and stored in advance for faster training.

3.2.1 Question Pre-processing

Questions are extracted and stored in a list from the JSON files provided by the VQA dataset. Since the model is fully based on the transformer, the questions are tokenized in the same form that BERT used during pre-training. BERT requires special tokens to be added before ([CLS]) and after ([SEP]) every sentence that informs the model about the beginning and end of the sentence. Then using the WordPiece tokenizer, words are split into either full forms or into word pieces. For example, running gets split into [‘run’, ‘##ing’]. Questions are made to a maximum of length 14 by using padding. If the question length is smaller than 14 then Pad tokens are added and if the length is greater than 14 then the question is truncated without losing any information. Attention masks are also generated to help the model differentiate between the pad tokens [PAD] so, the model does not attend to those padded features.

3.2.2 Answer pre-processing

Two different methods were used to extract the best answer. In the first method, we select the modal answer as the answer that occurred the most among the 10 labeled answers to be the official answer. In the second method, we reply to the multiple-choice answer which is machine-generated to get more consistent results. For soft label scores, the scores are generated based on the answer’s frequency in the set of 10 answers per question. The score is set according to the evaluation metric used in [17], so the model learns better and the accuracy is more aligned with the performance metric used. In all the above methods the answers are pre-processed in a similar way to the questions.

3.2.3 Image pre-processing

We used bottom-up features [32] for image feature extraction. The output feature of 2048 is generated for each image region using F-RCNN. The output from the above model is taken and the non-max suppression using IoU (Intersection of union) is applied to each object class. Only those object regions are selected for which confidence exceeds a given threshold. Figure 1 shows the output generated by F-RCNN.

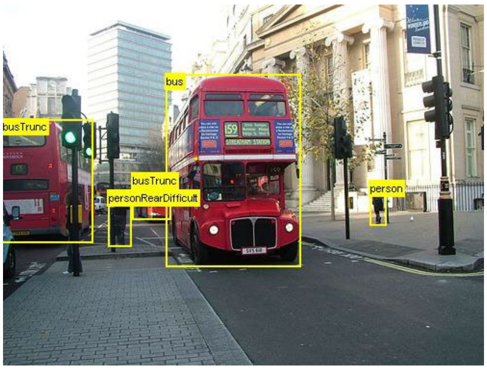


Figure 1: Faster-RCNN output

3.3 The Model Architecture

Transformers have recently become one of the most popular methods in NLP for working on unimodal tasks. However, they have also been used extensively in multi-modal tasks like VQA and image captioning. Many papers [23], [25], [26] propose using the encoder architecture of BERT for the entire VQA model. Inspired by the BERT's success in learning the language model, the proposed model has been implemented fully using transformer - BERT. The model uses the encoder-decoder strategy of the transformer introduced in [33]. The proposed TRANS-VQA model has two important components: encoder and decoder. The BERT has been used as an encoder to extricate the question features, as the BERT has already been trained for extracting language features. Additionally, the decoder's design is predicated on the BERT architecture, which collaborates to handle the image and question characteristics. The decoder attempts to extract rich image features that are directed by the question features after accepting the question features (extracted by the BERT-based encoder) and the image features (extracted by F-RCNN). Essentially, the question is fed into the BERT-based decoder, which then handles the question-guided image features. Figure 2 is an illustration of the suggested TRANS-VQA model's framework.

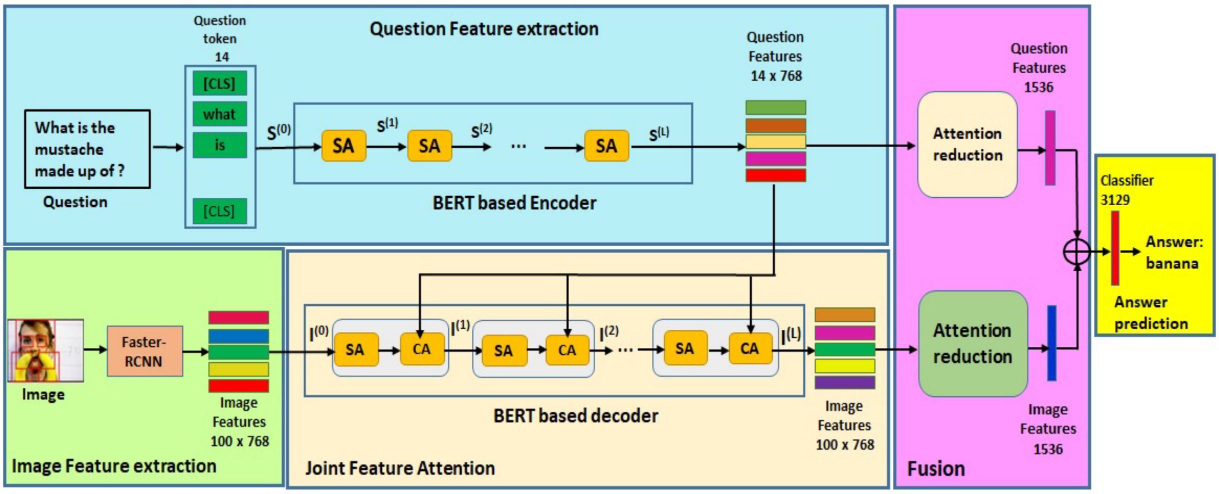


Figure 2: The TRANS-VQA model and its components

The architecture contains five modules: Question feature extraction, Image feature extraction, Joint feature representation, Multi-modal fusion, and finally, answer prediction Module.

Question feature extraction:

As discussed in section 3.2.1 the questions are extracted using BERT initialised with pre-trained weights.

Image feature extraction:

We make use of Bottom-Up (F-RCNN) features as discussed in section 3.2.2 with adaptive K where K is the number of regions extracted from the image. Due to the variable number of image features, we apply padding to each image so that the number of features is 100 and the corresponding mask is also used by the model to prevent attending to the padded regions.

Joint Feature Attention Module:

The joint feature attention has been performed by the decoder. The decoder contains many attention layers cascaded in depth. The decoder is designed based on the scaled dot product using transformers. The queries and the values with dimension d_{key} and d_{val} respectively are fed as an input to the decoder. Where $d_{val} = d_{key} = d$ set to the same dimension. For the VQA problem, the image features $I \in R^{mXd_I}$ are considered to be the queries, where d_I is the feature dimension, and $m \in [10,100]$ denotes the number of objects (for adaptive K). The keys and values are the question features $S \in R^{nXd_s}$ where d_s is the embedding size and $n \in [1,14]$ is the sequence length. To obtain the attention weights on the values first, the dot product of the query q with all the keys K is computed. Then each term is divided by $\sqrt{d}$ and finally apply a softmax function. This is given by:

$$f = A(q, K, V) = softmax\left(\frac{q\,K^T}{\sqrt{d}}\right) V \tag{2}$$

This is done on each head in multi-head attention, given by:

$$f = MultiHeadAtten(q, K, V) = (head_1\,, head_2, \dots \dots.., head_h)W^0 \tag{3}$$

$$head_j = Attention(qW_j^Q\,, KW_j^K\,, VW_j^V\,) \tag{4}$$

Where $W_J^Q, W_J^k, W_J^v \in R^{dXd_h}$ are the projection matrices for the jth head, and $W_0 \in R^{h*d^{hXd}}$. For each head, the dimension d_h is given by d_h=d/h where h denotes the number of heads. The Self-Attention (SA) unit shown in Figure 3 takes only one input S. Here, the job of the multi-head attention layer is to learn the pairwise interaction between the paired sample <*Si,Sj*> within S and generate the attended features $z \in R^{mXd}$ by weighted summation of all instances in S. The output feature generated by the multi-head attention layer is given to the feed-forward layer which and further transforms them through two 2-layer MLPs with GELU [34] activation in between given by:

$$FFN(I) = GELU(I\,W_1 + b_1)W_2 + b_2 \tag{5}$$

The input and out dimensions are the same as *d* while the inner layer has dimensionality, $d_{ff} = 4 * d$. Moreover, to facilitate optimization, residual connection [35] followed by layer normalization [36] is applied to the outputs of the two layers. The Co- Attention unit (CA) shown in Figure 3, takes two input features: I (Image features) and S (Question features) where question feature S directs the attention learning for image features I. The CA models the pairwise interaction between each paired sample <*Si,Ii*> from S and I, respectively. As shown in Figure 4, the decoder contains an SA unit followed by a CA unit.

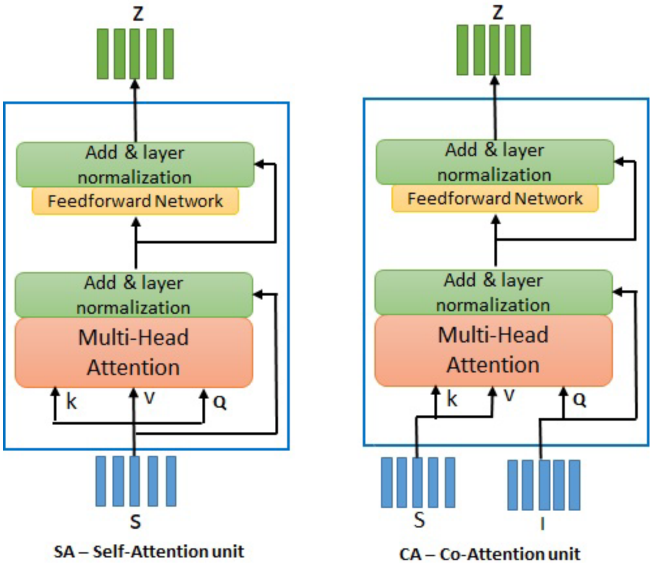


Figure 3: SA (Self-Attention) and CA (Co-attention) units

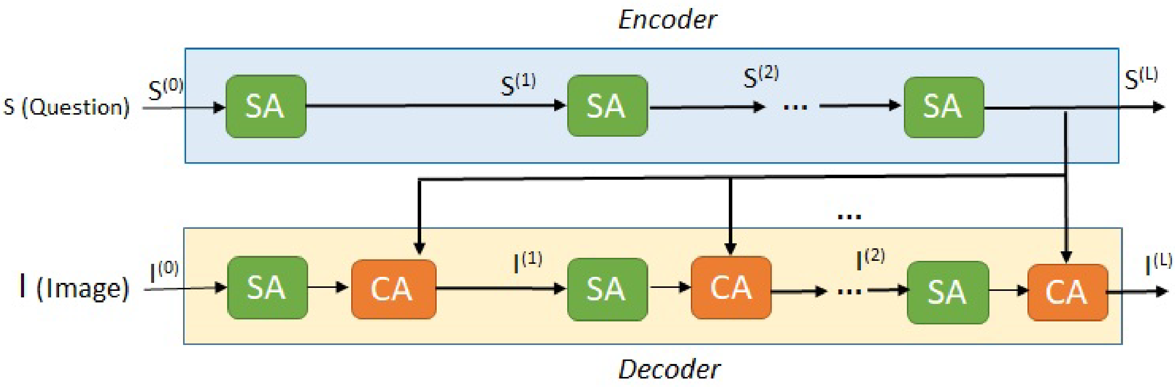


Figure 4: Transformer based Encoder-Decoder architecture

The decoder performs attention learning by passing the input features S (question) and I (image) through the decoder consisting of L number of attention layers cascaded in depth. We denote the input features as $I^{(l-1)}$ and S respectively, where S is the question features from the last layer of the encoder, which are further fed to the next layer of the decoder.

Multi-Modal fusion:

After co-attending to the question and image feature vectors, the two feature vectors need to be combined (or fused) together such that the output can be used to predict an answer. Most implementations go for simple element-wise multiplication, addition, or concatenation of the features to fuse them. We used the attention reduction model discussed below to combine the two features.

Attention reduction model:

The encoded question features and the decoded co-attended image features are flattened using the attention reduction model. After this stage, the output image, $I(L) = \{I_1(L), I_2(L), \dots\dots.., I_m(L)\} \in R^{mXd}$ and question features $S(L) = \{S_1(L), S_2(L), \dots\dots.., S_n(L)\} \in R^{nXd}$ contain contextual information about the attention weights over the image regions and question words. The attention reduction module contains a two-layer MLP for I(L) and S(L) to obtain their rich attended feature $\tilde{S}$ and $\tilde{I}$. Given I(L) as an input, the attended feature $\tilde{S}$ is obtained as follows:

$$\alpha = Softmax(MLP(I^L)) \tag{6}$$

$$\tilde{I} = \sum_{i=1}^{m} \left(MLP(\alpha_i I_i^{(L)})\right) \tag{7}$$

Where $\alpha = [\alpha_1, \alpha_2, \alpha_3 \dots, \alpha_m] \in R^m$ are the learned attention weights. The linear multi-modal fusion function is applied to the $\tilde{S}$ and $\tilde{I}$ as follows:

$$z = Layernorm(W_I^T \tilde{I} + W_S^T S) \tag{8}$$

Where d_z is the common dimensionality for the combined feature and $W_I, W_S \in R^{dXd_z}$ are two linear projection matrices. The use of LayerNorm here is to stabilize training.

Answer prediction Module:

The answer prediction module is a simple classifier. Once we combine the attended question feature vectors and image feature vectors, the result of the Fusion module is given as an input to a multi-layer perception (MLP) followed by a softmax function which then generates a vector of probabilities for probable answers from the entire answer set. Here, the VQA is treated as a multiple-label classification task. The model tries to predict the likelihood of one or more answers being applicable to a given question. Each question in a training set in VQA v2.0 is associated with more than one answer and is assigned a soft score in [0,1]. We create the output vocabulary containing all the answers appearing a minimum 8 times in the training set, giving 3129 candidate answers. The multi-label classifier uses sigmoid activation in the last layer to normalize the final scores between 0 to 1 which is followed by a binary cross-entropy loss, although we use soft target scores. The final step is a logistic regression that predicts the correctness of each candidate answer. The Loss function L is given as:

$$L = -\sum_i^M \sum_j^N S_{ij} \log(\hat{S}_{ij}) - (1 - S_{ij}) \log(1 - \hat{S}_{ij}) \tag{9}$$

Where the variables *i* and *j* denote the indices for *M* number of training questions and *N* number of candidate answers respectively. The ground-truth scores *S* are the aforementioned soft accuracies of ground-truth answers, generated as per the evaluation metric used by VQA.

4. Results

4.1 Experimental setup

The model was developed using TensorFlow and Keras. The final model was run on a machine learning server with 32GB NVIDIA GPU and 16-core CPU for training. The model was validated on the validation set after every epoch using the evaluation metric specified. Note that, when the model was evaluated on validation set, we used only the training split for training the model. Whereas, for evaluating the results on the test-dev and test-standard splits, we used the train split, the validation splits along with the subset of VQA samples from Visual Genome [37] for training.

Various ablation studies were performed on the proposed model by changing various hyperparameters and finally, we concluded the best hyperparameters to be used. All models were trained for 13 epochs. For optimizing hyperparameters the Adam optimizer with a learning rate of 1e-4 and β_1=0.9 and β_2=0.98 was used. The learning rate was set to $min(2.5te^{-5}, 1e^{-4})$, where *t* is the current epoch number starting from one. After 10 epochs, the learning rate is decayed by 1/5 every 2 epochs. Our best model has 24 layers, hidden size 1024 trained with batch size 128.

4.2 Evaluation Metric

The authors of the VQA dataset specified an evaluation metric considering the variability in phrasing the answers by different humans. The accuracy for a given answer was specified as:

$$\boldsymbol{Acc(ans) = min\left\{\frac{\#humans\ that\ said\ ans}{3}, 1\right\}} \qquad (10)$$

They took a subset of 9 answers from the 10 total answers and after calculating the accuracy using the formula given in eq (10), these accuracies are then averaged to give the actual accuracy for a given answer. Before evaluation, all the answers are pre-processed. All the characters are converted to lowercase, all periods except decimal points are removed, number words are converted to digits, all articles are removed and all punctuations are replaced with spaces. Apostrophes and colons are not removed since removing apostrophes may change the meaning of a word and colons often refer to time. These processing steps are done for actual answers also.

4.3. Results and discussion

All the proposed models were trained on the VQAv2 dataset and tested on the validation set of VQAv2 dataset discussed in section 3.1. The results of open-ended answers for the **Validation set** of VQA v2.0 have been compared with previous models in Table 1.

Table 1: Accuracies (in %) on validation set of VQAv2 dataset

Model	**Yes-No**	**Num**	**Other**	**All**
HieCoAtt [13]	71.80	36.53	46.25	54.57
Top-Down Attention [32]	80.3	42.87	55.81	63.2
Counter [28]	-	49.36	-	65.42
BAN [38]	-	-	-	65.81
DFAF [39]	-	-	-	66.66
MRA-NET [29]	**-**	**-**	**-**	66.08
TRANS-VQA (BERT-BASE)	**84.87**	**48.31**	**58.66**	**67.15**

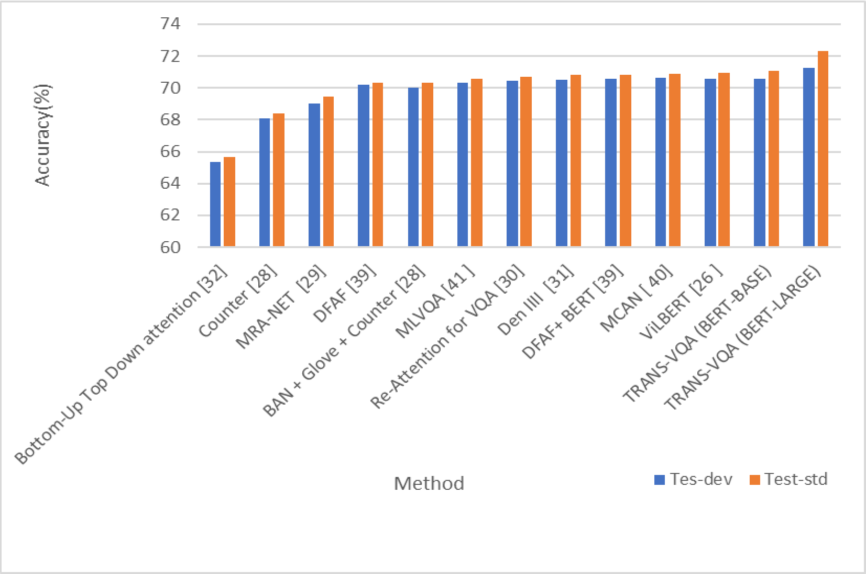
TRANS-VQA (BERT-LARGE)	**85.01**	**48.88**	**60.79**	**68.21**

As shown in Table 1 our final model reaches a validation accuracy of around **68.21%** which shows improvement over all previous models on VQA. A comparison of the proposed model with previous models evaluated on the VQA test-dev and test-std set is given in Table 2. Note that, the proposed TRANS-VQA model is trained on the Train set + Validation set + Visual Genome dataset and evaluated on the VQA test-dev and test-std set.

Table 2 shows that the proposed model outperforms all previous models on the VQA v2 test-dev and test-std set. We can infer that BERT as a language model's ability to understand meaning allows it to gain an advantage over traditional Glove-based methods easily and gives better performance than most models in answering Yes/No questions. However, it fails to offer better performance in Number questions, which most models struggle against. BAN+Glove+Counter performs better at counting tasks which was the main focus of the counter [28] paper. By integrating counter into the proposed model, number counting prediction can be further improved. We managed to get an increase of around 2% on the test-std dataset against similar SOTAs, which suggests our model reaches the same understanding as a human. Model ensembeling and retraining can further improve our current performance. The comparison of the proposed model with previous state-of-the-art models on test-dev and test-std accuracies on the VQAv2 dataset has been shown in Figure 5.

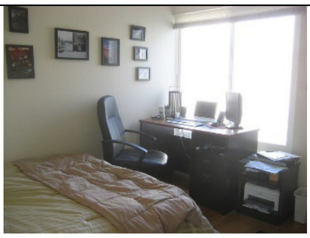
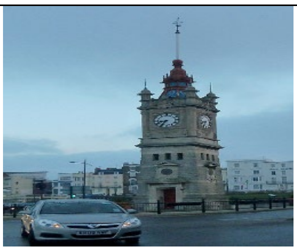
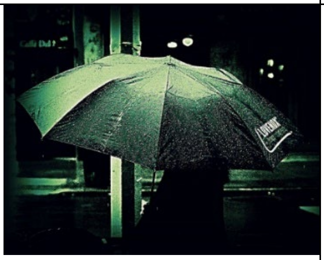
Table 2: accuracy of the final suggested model on the test-standard and test-dev datasets for VQA v2.0. The Visual Genome dataset is used to train the models, together with training and validation splits.

Method	**Test-development**				**Test-standard**
	Yes/No	**Number**	**Other**	**All**	**All**
Top Down attention [32]	81.82	44.21	56.05	65.32	65.67
Counter [28]	83.14	51.62	58.97	68.09	68.41
MRA-NET [29]	85.58	48.92	59.46	69.02	69.46
DFAF [39]	86.09	53.32	60.49	70.22	70.34
BAN & Glove & Counter [28]	85.42	**54.04**	60.52	70.04	70.35
MLVQA [41]	86.64	51.90	60.53	70.30	70.57
Re-Attention for VQA [30]	87.00	53.06	60.19	70.43	70.72
Den IIII [31]	86.30	50.9	61.5	70.50	70.8
DFAF+ BERT [39]	86.73	52.92	61.04	70.59	70.81
MCAN [40]	86.82	53.26	60.72	70.63	70.90
ViLBERT [26]	-	-	-	70.55	70.92
TRANS-VQA (BERT-BASE)	86.96	52.87	60.53	70.59	71.05
TRANS-VQA (BERT-LARGE)	**87.04**	52.92	**61.05**	**71.23**	**72.31**



.

Figure 5: Graph of the accuracy of the proposed model and previous models

 Image ID: 262162 Q: Which room is this? A: bedroom ✓ Ground truth : bedroom	 Image ID: 393277 Q: What is in front of the tower? A: car ✓ Ground truth : car	 Image ID: 262229 Q: How many girls are playing? A: 4 ✓ Ground truth : 4
 Image ID: 153994 Q:What does the white writing on the umbrella say? A: umbrella ✗ Ground truth : lovenox	 Image ID: 22935 Q: What is floating? A: soccer ball ✓ Ground truth : Soccer ball	 Image ID: 25668 Q: Are this people real? A: yes ✓ Ground truth : yes

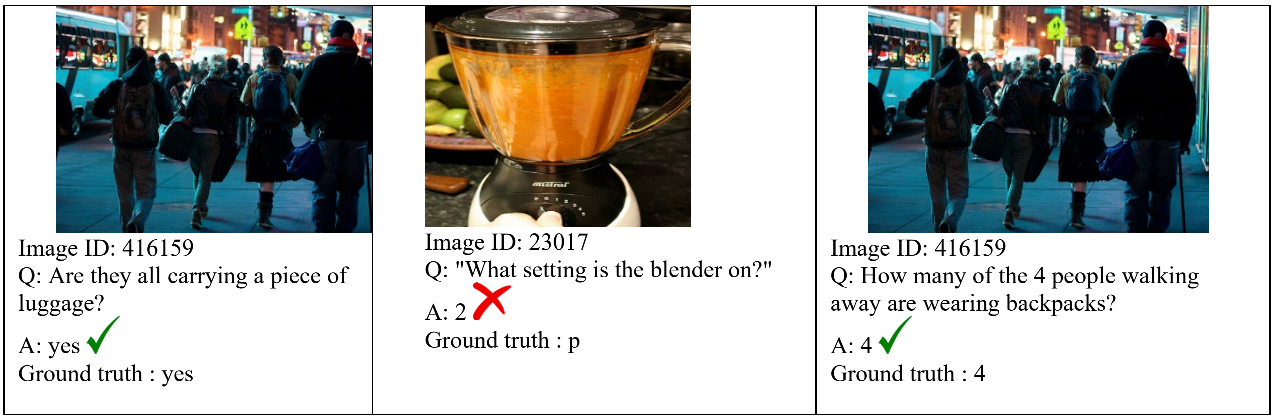


Image ID: 416159
Q: Are they all carrying a piece of luggage?
A: yes ✓
Ground truth : yes

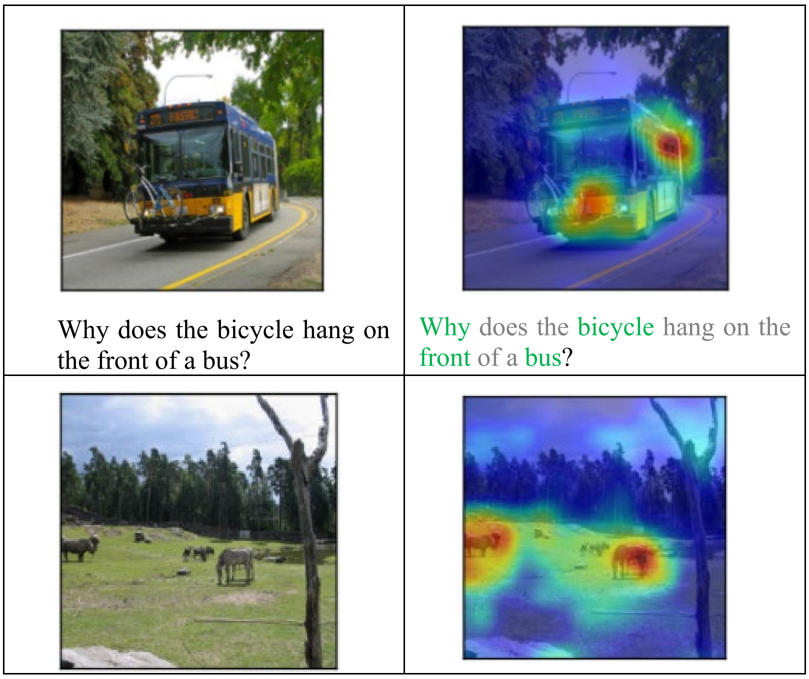
Image ID: 23017
Q: "What setting is the blender on?"
A: 2 ✗
Ground truth : p

Image ID: 416159
Q: How many of the 4 people walking away are wearing backpacks?
A: 4 ✓
Ground truth : 4

Figure 6. Model-generated answers and ground truth on validation images of VQA dataset

A selection of sample photos from the validation set along with our best model's projected responses for each question are displayed in Figure 6. It is evident that the model correctly predicts the answer the majority of the time. The algorithm occasionally fails to anticipate the right response, for example, when the image is unclear or when the answer is obscured by non-human-readable language in the image. Also, the questions whose answers are not directly present in the image and require outside knowledge have not been projected properly.

Figure 7 shows question attention as well as visual attention maps obtained for a few of the sample images by the TRANS-VQA model and gives insight into how visual attention is drawn to the important parts of an image.



Why does the bicycle hang on the front of a bus?

Why does the bicycle hang on the front of a bus?

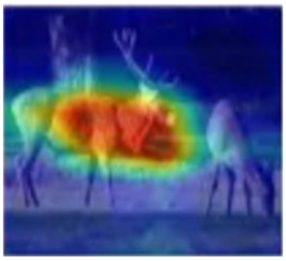
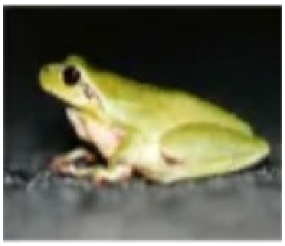
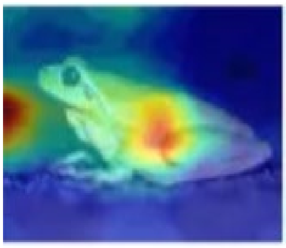
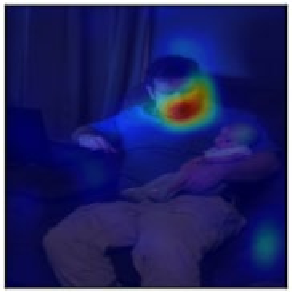
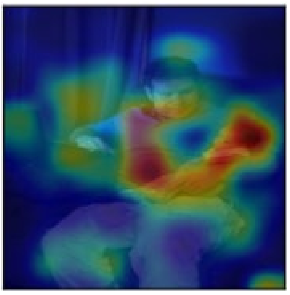
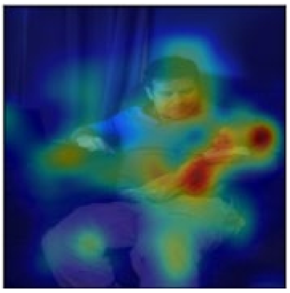
Are these animals in the wild or civilized environment?	Are these animals in the wild or civilized environment?
 How many deer are there?	 How many deer are there?
 What color is the frog?	 What color is the frog?

Figure 7. Question attention and corresponding Visual attention maps obtained for a few example (image, question) pairs of the VQA dataset. Green highlighted text shows important and attended question words.

Original Image	**Layer 6 attention**	**Layer 8 attention**	**Layer 12 attention**
 What is there in a man's hand?			

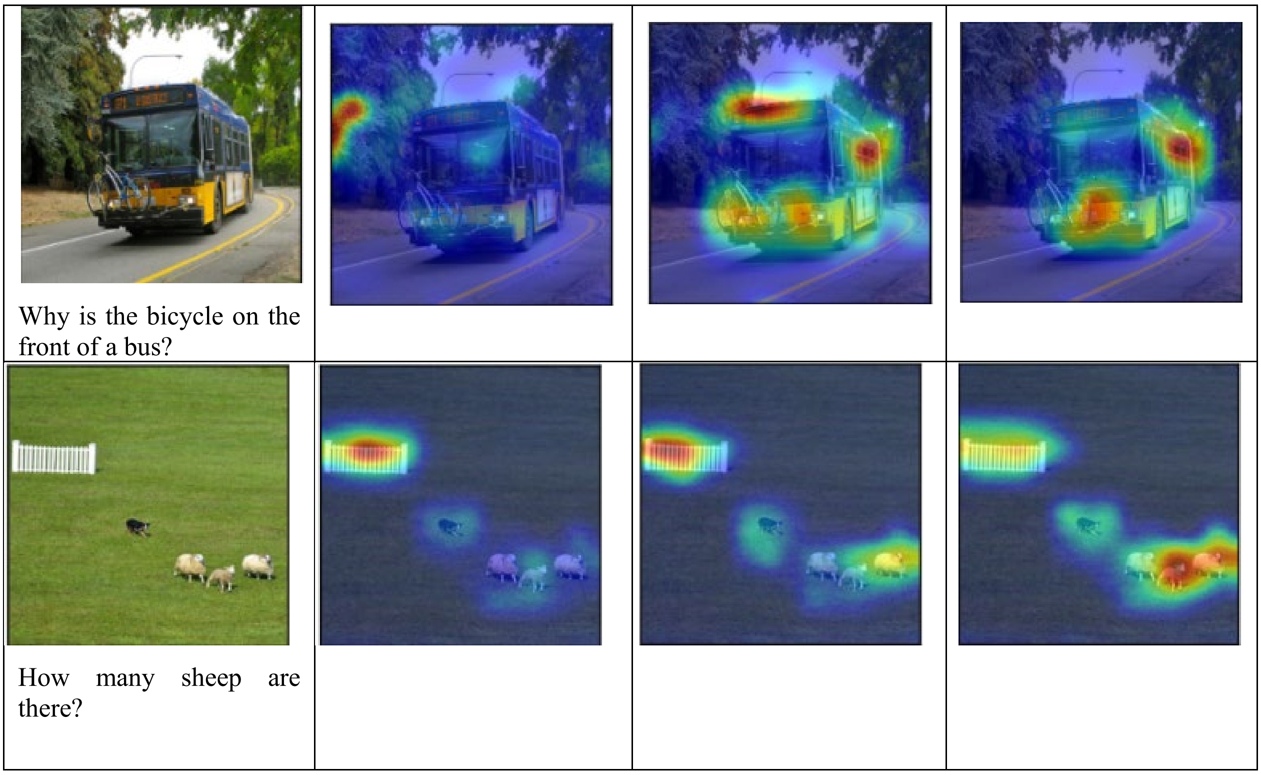

Why is the bicycle on the front of a bus?
How many sheep are there?

Figure 8: Visual attention maps at layers 6,8 and 12 of BERT-based decode for a given Image, question pair

To examine the impact of adding layers to the encoder-decoder of the suggested TRANS-VQA model on visual attention, we have analyzed the visual attention map at different layers of the decoder i.e. after layers 6, 8, and 12 shown in Figure 8. It has been observed that at each increasing layer, visual attention improves. This immediately confirms our conclusion that the accuracy of the model grows with the number of layers in the encoder and decoder.

4.4 Ablation Study

We undertake a series of ablations to test the impact of different hyperparameters on the performance of the suggested model. A single or two alterations to the reference model are taken into account in each of the ablation experiments displayed in Table 3. The outcomes confirm the patterns found with the individual ablations.

Increasing the number of layers in the transformer increases the accuracy of the model. Also, using the adaptive K image features provides better results than the fixed image features. We can see that increasing the BERT size also improves performance, though the compact models do offer competitive performance. Using BERT with pre-trained features is also essential in improving the model's learning capacity. Interestingly, performance seems to drop when decreasing or increasing the batch size suggesting that matching the batch size with the per attention-head key dimension in the transformer architecture is what contributes to a higher score. Increasing the question length also does not seem to improve performance but rather hurts it. We experimented with different activation functions but because BERT was trained with the GELU activation function, it offers better performance during fine-tuning. These experiments reveal our best model to have the following configuration: 24 layers, approximate GELU activations, 14 question length input, bottom-up attention features with adaptive K, pre-trained BERT-Large model for encoder, and batch size of 128.

Table 3: A single network's ablations are assessed using data from the VQA v2 validation set. We assess different iterations of our top "reference" model (shown in the first row). For single alterations, we train each version using a constant random seed.

Reference model	**VQA v2.0 validation set**

	Yes - No	Num	Other	All
	84.17	48.66	58.35	66.78
Number of encoder/decoder layers (ref: 8 layers)				
2 layers	83.02	45.96	57.51	65.58
4 layers	83.64	46.65	58.06	66.18
6 layers	83.8	47.69	58.39	66.54
Hidden activation (ref: using approximate GELU activation function)				
Using GELU activation function	84.16	48.48	58.34	66.75
Using ReLU activation function	83.86	47.83	58.36	66.56
Using SiLU activation function	83.47	48.46	57.89	66.26
Question length (ref: 14 words per question)				
28 words per question	84.12	48.24	58.4	66.73
Image features (ref: bottom-up attention features, adaptive K)				
Bottom-up attention features, K=36	83.74	47.38	57.93	66.25
BERT models (ref. using BERT-Medium)				
Using BERT-Small	83.64	46.65	58.06	66.18
Using BERT-Base	84.87	48.31	58.66	67.15
Using BERT – Large	85.01	48.42	58.67	**68.21**
BERT pretraining (ref. pretrained weights)				
Randomly initialized weights	83.42	48.01	57.98	66.23
Mini-batch size (ref: 64 training questions)				
32 training questions	83.71	48.13	57.42	66.08
128 training questions	83.79	47.16	58.25	66.39

Table 4: Effect of Number of Layers in encoder/decoder on VQAv2 Validation set.

No. of layers in Encoder/Decoder	Yes/No	Number	Others	All
L=2	83.02	45.96	57.51	65.58
L=4	83.64	46.65	58.06	66.18
L=6	83.8	47.69	58.39	66.54
L=8	84.17	48.66	58.35	66.78

L=12	84.23	48.69	58.34	67.15
L=24	85.01	48.88	60.79	68.21

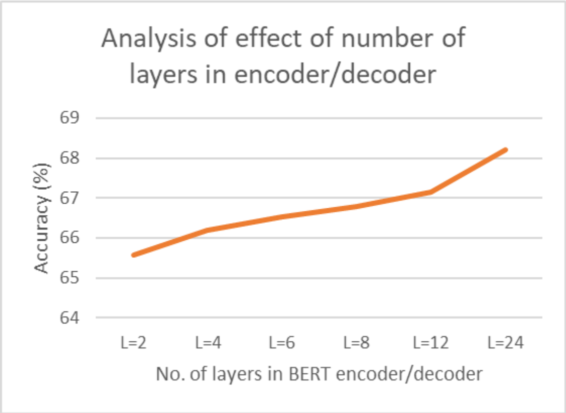
Table 5: Effect of various activation functions on VQA-v2 Validation

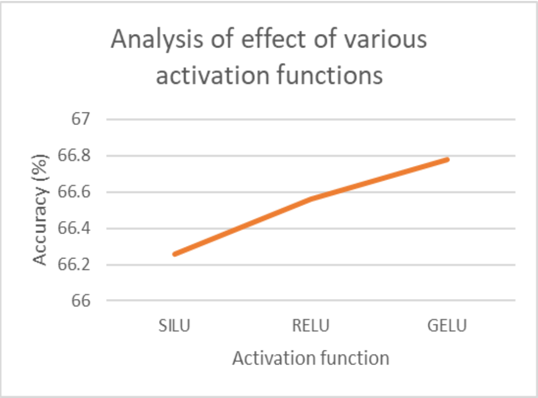
Activation Function	**Yes/No**	**Num**	**Others**	**All**
GELU	84.16	48.48	58.34	66.78
RELU	83.86	47.83	58.36	66.56
SILU	83.47	48.46	57.89	66.26

Table 6: Ablation studies of various BERT models on VQA-v2 Validation

BERT Models	**Layer, Hidden nodes**	**Yes/No**	**Num**	**Others**	**All**
BERT Small	L=4 , H=512	83.64	46.65	58.06	66.18
BERT Medium	L=8, H=512	84.17	48.66	58.35	66.78
BERT Base	L=12, H=768	84.87	48.31	58.66	67.15
BERT Large	L=24, H=1024	85.01	48.42	58.67	68.21

Table 3, Table 4, and Table 5 show the effect of varying numbers of layers in the encode/decoder, Activation functions, and BERT models respectively on the model performance. We derive the conclusion that the model accuracy grows with the number of layers in the encoder/decoder. Also, GELU activation outperforms the other two activations. Although BERT large gives the best accuracy, training the BERT large model requires more GPU and memory usage. The same results have been shown graphically in Figure 9.





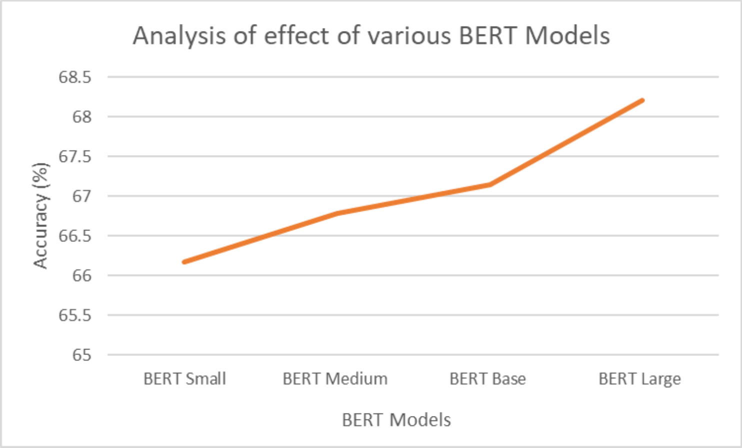


Figure 9: Analysis of the effect of various hyperparameters on model accuracy

5. Conclusion and future work

We proposed a novel fully transformer-based visual question-answering model using question-guided visual attention. The implemented model is based on the encoder-decoder strategy used in transformers. Extensive ablation was performed on the reference model by changing various hyperparameters to observe the effect of each hyperparameter on the reference model. After performing ablation, we came up with our best model configuration as 24 layers in encoder and decoder, approximate GELU activations, 14 question length input, bottom-up attention features with adaptive K, pre-trained BERT-Large model for encoder, and batch size of 128. The BERT–based model reaches **72.31%** accuracy and outperforms all SoTA. Based on our experiments, we conclude that Transformer architectures have a large scope in solving the Visual Question Answering task. Further understanding of how the decoder architecture helps image co-attention will allow us to improve it. Future work may include combining local and global image features to improve the extraction of valuable features. Also, integrating outside knowledge into the model will help the model to answer open-domain questions that cannot be answered by the proposed model.

References

[1] Malinowski M, Rohrbach M, and Fritz M. Ask: Your Neurons: A Neural-based Approach to Answering Questions about Images. In Proc. IEEE Int. Conf. Comp. Vis., 2015.

[2] Gao H, Mao J, Zhou J, Huang Z, Wang L, Xu W. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering. In Proc. Advances in Neural Inf. Process. Syst., Volume 2 , pages 2296-2304, 2015.

[3] Zhou B, Tian Y, Sukhbaatar S, Szlam A, Fergus R. Simple Baseline for Visual Question Answering. arX-iv Preprint. arXiv preprint, https://doi.org/10.48550/arXiv.1512.02167. 2015.

[4] Yang Z., He X., Gao J, Deng L and Smola A. Stacked Attention Networks for Image Question Answering," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21-29, 2016.

[5] Noh H, Seo PH, Han B. Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 30-38), 2016.

[6] Chen K, Wang J, Chen LC, Gao H, Xu W, Nevatia R. Abc-cnn: An attention-based convolutional neural network for visual question answering. arXiv preprint arXiv:1511.05960. 2015.

[7] Ilievski I, Yan S, and Feng J. A Focused Dynamic Attention model for visual question answering." [Online]. Available: https://arxiv.org/abs/-1604.01485. 2016

[8] Kafle K. and Kanan C. Answer-Type Prediction for Visual Question Answering, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4976-4984, 2016.

[9] Nguyen D, and Okatani T. Improved Fusion of Visual and Language Representations by Dense Symmetric Co-attention for Visual Question Answering, IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6087-6096, 2018.

[10] Shih K. J, Singh S and Hoiem D. Where to Look: Focus Regions for Visual Question Answering, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4613-4621, 2016.

[11] Zhang P, Goyal Y, Summers-Stay D, Batra D, and Parikh D. Yin and Yang: Balancing and Answering Binary Visual Questions, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5014-5022, 2016.

[12] Lobry S, Marcos D, Murray J, Tuia D. RSVQA: Visual Question Answering from Remote Sensing Data, IEEE Transactions on Geoscience and Remote Sensing, 2020.

[13] Lu J., Yang J., Batra D., and Parikh D. Hierarchical question image co-attention for visual question answering, In Proc. NIPS, 2016, pp. 289_297. 2016.

[14] Yang C, Jiang M, Jiang B, Zhou W and Li K. Co-Attention Network With Question Type for Visual Question Answering, in IEEE Access, vol. 7, pp. 40771-40781,201, 2019.

[15] Chowdhury I, Nguyen K, Fookes C, Sridharan S. A cascaded long short-term memory (LSTM) driven generic visual question answering (VQA). In 2017 IEEE International Conference on Image Processing (ICIP), pp. 1842-1846, IEEE, 2017.

[16] Ziaeefard M, and Lecu, F. Towards Knowledge-Augmented Visual Question Answering. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 1863–1873). International Committee on Computational Linguistics, 2020.

[17] Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D. VQA: visual question answering Proceedings of the IEEE International Conference on Computer Vision, pp. 2425–2433, doi: 10.1109/ICCV.2015.279, 2015.

[18] Lu J, Batra D, Parikh D, and Lee S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tas Maryam ks. In *NeurIPS* (pp. 13-23),2019

[19] Nguyen BX, Do T, Tran H, Tjiputra E, Tran QD, Nguyen A. Coarse-to-Fine Reasoning for Visual Question Answering, in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 2022 pp. 4557-4565, doi: 10.1109/CVPRW56347.2022.00502, 2022.

[20] Li X, Yin X, Li C, Zhang P, Hu X, Zhang L, Wang L, Hu H, Dong L, Wei F, Choi Y. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks, In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16 2020 (pp. 121-137). 10.1007/978-3-030-58577-8_8, 2020.

[21] Yang Z, He X, Gao J, Deng L, Smola A. Stacked Attention Networks for Image Question Answering, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21-29, 2016 doi: 10.1109/CVPR, 2016.

[22] Gao L, Cao L, Xu X, Shao J, Song J. Question-Led object attention for visual question answering, in Neuro computing, Volume 391, Pages 227-233, ISSN 0925-2312,2020.

[23] Nguyen DK and Okatani T. Improved Fusion of Visual and Language Representations by Dense Symmetric Co-attention for Visual Question Answering, IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6087-6096, doi: 10.1109/CVPR.2018.00637, 2018.

[24] Rahman T, Chou SH, Sigal L, Carenini G. An Improved Attention for Visual Question Answering, IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1653-1662, doi: 10.1109/CVPRW53098.2021.00181, 2021.

[25] Nguyen BX, Do T, Tran H, Tjiputra E, Tran QD, Nguyen A. Coarse-to-Fine Reasoning for Visual Question Answering, in IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 2022 pp.4557-4565, doi: 10.1109/CVPRW56347.2022.00502, 2022.

[26] Lu J, Batra, D, Parikh D & Lee S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. B. Fox & R. Garnett (eds.), NeurIPS (p./pp. 13-23), 2019.

[27] Zhong H, Chen J, Shen C, Zhang H, Huang J, Hua XS. Self-Adaptive Neural Module Transformer for Visual Question Answering, in IEEE Transactions on Multimedia, vol. 23, pp. 1264-1273, doi: 10.1109/TMM.2020.2995278, 2021.

[28] Zhang Y, Hare J, Prügel-Bennett. Learning to Count Objects in Natural Images for Visual Question Answering. In International Conference on Learning Representations, 2018.

[29] Peng L, Yang Y, Wang Z, Huang Z, Shen HT. MRA-Net: Improving VQA Via Multi-Modal Relation Attention Network, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 1, pp. 318-329, doi: 10.1109/TPAMI.2020.3004830, 2022.

[30] Guo W, Zhang Y, Yang J, Yuan X. Re-Attention for Visual Question Answering, in IEEE Transactions on Image Processing, vol. 30, pp. 6730-6743, doi: 10.1109/TIP.2021.3097180,2021.

[31] Liu F, Liu J, Fang Z, Hong R, Lu H. Visual Question Answering with Dense Inter- and Intra-Modality Interactions, in IEEE Transactions on Multimedia, vol. 23, pp. 3518-3529, doi: 10.1109/TMM.2020.3026892, 2021.

[32] Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6077-6086, doi: 10.1109/CVPR.2018.00636, 2018.

[33] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010, 2017.

[34] Hendrycks D, and Gimpel K. Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units. ArXiv, abs/1606.08415, 2016.

[35] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 770-778. 10.1109/CVPR.2016.90, 2016.

[36] Ba JL, Kiros JR, Hinton GE. Layer Normalization. ArXiv, abs/1607.06450, 2016.

[37] Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li LJ, Shamma DA, Bernstein MS. Visual genome: Connecting language and vision using crowdsourced dense image annotations, International Journal of Computer Vision, 123(1):32–73, https://doi.org/10.1007/s11263-016-0981-7, 2017.

[38] Fang P, Zhou J, Roy SK, Petersson L, Harandi M. Bilinear Attention Networks for Person Retrieval," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8029-8038, doi: 10.1109/ICCV.2019.00812, 2019.

[39] Gao P, Jiang Z, You H, Lu P, Hoi SC, Wang X, Li H. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6639–6648, 2019.

[40] Yu Z, Yu J, Cui Y, Tao D, Tian Q. Deep modular co-attention networks for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6281–6290,2019.

[41] Ma J, Liu J, Lin Q, Wu B, Wang Y, You Y. Multitask Learning for Visual Question Answering, in IEEE Transactions on Neural Networks and Learning Systems, doi: 10s.1109/TNNLS.2021.3105284, 2021.