

BERT based Hierarchical Alternating Co-Attention Visual Question Answering using Bottom-Up Features

Dipali Koshti¹, Dr. Ashutosh Gupta², Dr. Mukesh Kalla³

Submitted: 18/09/2022

Accepted: 21/12/2022

Abstract: Answering a question from a given visual image is a very well-known vision language task where the machine is given a pair of an image and a related question and the task is to generate the natural language answer. Humans can easily relate image content with a given question and reason about how to generate an answer. But automation of this task is challenging as it involves many computer vision and NLP tasks. Most of the literature focus on a novel attention mechanism for joining image and question features ignoring the importance of improving the question feature extraction module. Transformers have changed the way spatial and temporal data is processed. This paper exploits the power of Bidirectional Encoder Representation from Transformer (BERT) as a powerful question feature extractor for the VQA model. A novel method of extracting question features by combining output features from four consecutive encoders of BERT has been proposed. This is from the fact that each encoder layer of the transformer attends to features from the word to a phrase and ultimately to a sentence-level representation. A novel BERT-based hierarchical alternating co-attention VQA using the Bottom-up features model has been proposed. Our model is evaluated on the publicly available benchmark dataset VQA v2.0 and experimental results prove that the model improves upon two baseline models by 9.37% and 0.74% respectively.

Keywords: VQA, Visual-Question Answering, BERT-based VQA, Hierarchical VQA, Image Question answering.

1. Introduction

Answering an Image - question (VQA) is a much-known vision-language task in which the machine has to answer the question asked from an image [1]. Humans can easily reason over the content of an image, analyze the question, relate the question with the image content and, answer the given question; but for a machine, this entire task is challenging as the task involves two different modalities – vision and language. It involves vision analysis as image content needs to be interpreted and learned. It also involves language analysis as questions need to be analyzed and meaningful semantic features need to be extracted. Thus, VQA is a blend of NLP and Computer vision. Examples of Image question answering are shown in Fig 1.



Fig.. 1. Examples of Image-question answering

Once image features and question features are extracted, they need to be combined together in order to understand the relation between them. VQA is challenging because it involves different vision-related tasks such as object detection, scene detection, object counting, color detection, Object segmentation, and much more. With the rapid developments in deep learning models, this task has become easy. Also, for language modeling various deep learning models such as RNN and LSTM are available. Any VQA task basically involves four stages shown in Fig. 2. 1) Extraction of Image Features 2) Extraction of Question Features 3) Joint comprehension of question and image features 4) Generation of answer.

The first stage involves extracting image features using various CNN models. Most of the literature extracts the image features using pre-trained VGGNet [1-4], ResNet [5-7], googleNet [8-11] CNN models. These models

¹ Department of Computer Science and Engineering, Sir Padampat Singhania University, Rajasthan, India.

dipali.koshti@spsu.ac.in

² Department of Computer Science and Engineering, Sir Padampat Singhania University, Rajasthan, India.

ashu.gupta@spsu.ac.in

³ Department of Computer Science and Engineering, Sir Padampat Singhania University, Rajasthan, India.
mukesh.kalla@spsu.ac.in

extract global features of the image. Most of the recent literature used F-RCNN [12-17] to extract object-level features. Object-level features provide a deeper understanding of the image than global features. Some literature combines both global and object-level features to take the advantage of both global and local features [18,19]. Combining both global and local image features certainly improve model performance.

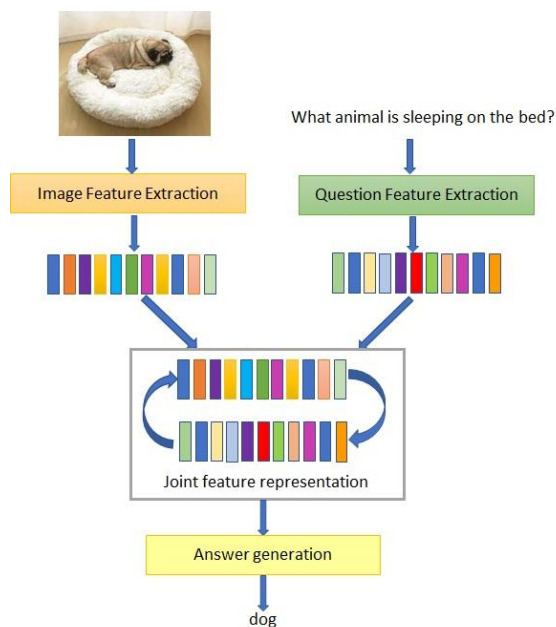


Fig. 2. Stages of Basic VQA model

The model performance can be improved by improving either image featurization or question featurization or by improving the joint representation phase. The proposed model improves the language model (question featurization) by employing transformers with hierarchical co-attention for question feature extraction. The model explores the power of BERT (Bi-directional Recurrent Transformers) as a powerful language feature extractor for visual question answering. A novel method of extracting hierarchical features from a question using BERT has been proposed.

2. Related Work

As VQA involves multiple modalities, it has attracted many researchers from Computer vision and NLP fields to work upon, and a variety of VQA models have been published in the past. Much of the literature uses visual attention mechanisms by dividing the image into regions and then attending to only important regions of the image. In [21] Kevin J et al. presented a VQA model using visual attention. The model selects only those visual regions that are interrelated to the question. For performing joint comprehension, the question features and the relevant image features are placed in a shared space and performed an inner product that assigns weighing to each region. The model performance is tested on the VQA 1.0 multiple-

choice dataset and achieved an accuracy of 62.43%. In [22] Ryan Burt et al. divided the image into separate objects and extracted the location and size of each object and these objects are then passed to the MLP. Authors used Gamma saliency to identify the most important objects in an image. The model was evaluated on the MNIST digit dataset. Zichao et al. [23], proposed a VQA model SAN (Stacked Attention Network) where semantic features of questions are used to search relevant regions of the image, and performed multi-step reasoning. Authors claimed that image-question answering performance can be improved by doing multi-step reasoning. The model was tested on COCO-QA and DAQUAR datasets with 58.75% and 29.3% accuracy respectively. The problem with these models is that they only use visual attention. The model performance can be improved by performing co-attention means performing both visual+ question attention.

In [23] Duy-Kein et al. presented a dense co-attention mechanism where he explored every possible interaction between the given question and an image by allowing each question word to guide the image regions and each image region to guide the question word. One attention map is generated for each word in a question. Similarly, one attention map is generated for each region in an image. The model was evaluated on VQA v1.0 and VQA v2.0 datasets with an accuracy of 66.66% and 67.0% on the Test-std set. In [24]. Lu J. et al. proposed a hierarchical approach to VQA where they analyzed the question at three different hierarchies: Phrase, word, and sentence level. They proposed two novel attention mechanisms: alternate and parallel co-attention mechanisms. The model was evaluated on VQA 1.0 dataset and they achieved an accuracy of 62.1%. A similar hierarchical approach is used in [25] with novel multi-step reasoning and adaptive fusion method. Here, the model performs textual and, visual attention through two steps. Also, adaptive reasoning assigns weights or importance to each of the word embedding (word, phrase, and sentence) adaptively.

Model performance can be improved by improving image featurization or question featurization module. Much of the earlier literature uses VGG net [1-4] for extracting image features. With the availability of computing resources, people have started using ResNet [5-7] for image feature extraction which is heavier than VGGnet but provides better and richer features than VGGNet. Now, most of the literature uses the F-RCNN bottom-up approach for extracting object-level image features [12-17]. In [13] authors proposed a VQA model that also integrates question type in a model to predict the answer. The model used F-RCNN for extracting object-level image features and Bi LSTM for question features. They proposed a novel co-attention mechanism where first self-attention on a question is performed in order to identify important words in a question. Then, this attended question is used to guide

the visual image and finds out important objects in an image. The model predicts the question type and feeds it to the MLP along with the attended image and question feature. In [26] Fukui et al. presented a novel joint representation technique called MCB for VQA where image and text representation is randomly projected to a higher dimensional space and then join both vectors efficiently by using the element-wise product in Fast Fourier Transform (FFT) space. In [27] authors have used a transformer as a language model and proposed a VQA model with Multiple Parallel Co-attention. They proved that transformers as a language model improve question feature extraction.

3. Theoretical Background

Transformers have revolutionized the field of language processing and image processing. Compared to traditional recurrent networks like RNN and LSTM, transformers have proven excellent in processing temporal information. Transformer works on the principle of self-attention. The attention function takes three inputs: Q, K, and V where K, Q, and V represent query, Key, and value respectively. A simple scaled dot product attention presented in [20] is given below.

$$Att(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (1)$$

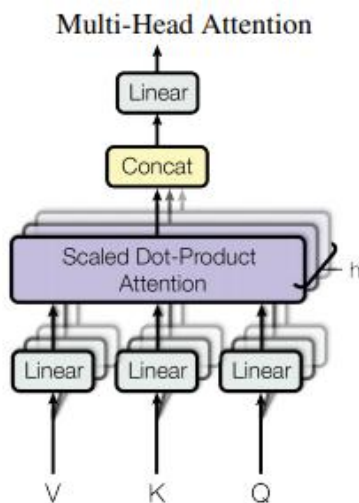


Fig. 3. Multi-head attention

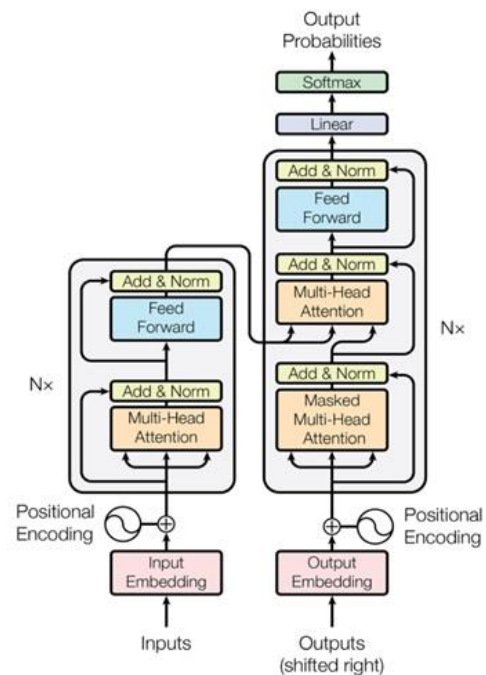


Fig.4. Architecture of Transformer

Fig. 3 shows multi-head attention mechanism. Here 8 different representations of Q, K and V with different random weights, go through the attention mechanism in parallel. Final results are aggregated and transformed onto expected output. Fig. 4 shows architecture of transformer. First, we have the encoder that contains two sub-layers: Multi – head attention layer and feed forward layer. On right hand side is the decoder that contains one more layer called masked multi head attention layer. In full architecture, we have such 6 stacks of encoder and decoders. Recently, many literatures used pre-trained transformer models. These models are pre-trained meaning they have been trained on an enormous corpus of words and sentences. Most popular one is Bidirectional Encoder Representation from Transformer model-BERT. BERT is basically an Encoder stack of transformer architecture. There are various versions of pre-trained BERT models available. We have used BERTBASE that has 12 layers in the Encoder stack.

4. Proposed Model

Given a pair of (I, S) where ‘I’ is the given image and ‘S’ is the question sentence in a natural language, answer ‘A’ is to be generated. Our model contains four major modules: 1) Image Featurization module 2) Question featurization module 3) Joint comprehension Module and 4) Answer generation module. The following section explains each module of the model in detail.

4.1 Image Feature Extraction

In order to extract meaningful features from a given image, Faster R-CNN is used [28]. F-RCNN allows the extraction of object-level features as opposed to the global grid

features extracted using the traditional CNNs as shown in Fig 5. Here, Faster R-CNN with ResNet-101 is used to locate various objects present in an image. Then non-maximum suppression is performed for each detected object class using an Intersection over Union (IoU) threshold. All the regions are compared with the confidence threshold. Only those regions for which the class detection probability exceeds this threshold are selected. For each such selected region, image feature vectors of dimension 2048 are generated as the mean-pooled convolutional feature from that region. The model used is pre-trained on Visual Genome data and then used to generate the features for the input images. The features generated per image are variable based on the fixed threshold used for object detections. Fig. 5 shows how image regions have been predicted by F-RCNN model. This process can be shown mathematically as follows. Given an image 'I', image feature vector 'v' is calculated as,

$$v = f(I) = \{v_1, v_2, \dots, v_k\}, v_i \in R^k \quad (2)$$

Where v_i the feature of i^{th} object and, function (*) indicates F-RCNN model. Here 'K' represents the number of object regions detected and 'k' represents the size of each object feature. Thus, extracted image feature vector is represented as $v \in R^{K \times k}$, $k = 2048$ and $K=100$ (Variable between 10 to 100)

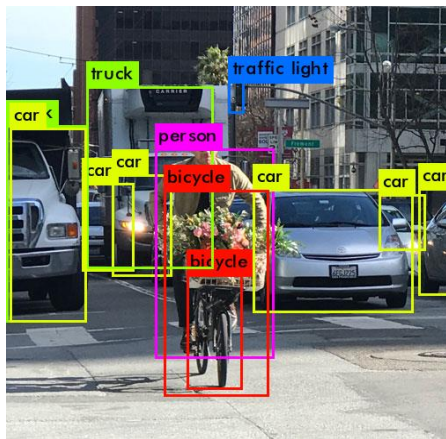


Fig. 5. Faster -RCNN Image

4.2 Question Feature Extraction

For extracting features at the word level from a given question, we used the pre-trained BERT model [26]. The question sentence is pre-processed before we input it to BERT. The BERT Tokenizer is used to tokenize the question. Input to the BERT model must start with "[CLS]" and end with "[SEP]" tokens so, [CLS] and [SEP] tokens were added accordingly to each question. WordPiece tokenizer was applied to split words into full forms or word pieces e.g. [surfing] is split to [surf, ##ing]. Questions are padded to a max length of 14. Attention

masks are generated to mask out [PAD] tokens. The output of the pre-trained BERT model is taken as contextual sentence embedding features. BERT has the capability of learning contextual relations between words in a text, unlike traditional word2vec implementations. Given a question sentence S with N number of words, a question is represented as,

$$S = \{s_1, s_2, s_3, \dots, s_N\} \quad (3)$$

First, we transform question S into lower dimension q using BERT

$$q = BERT(S) = \{q_1, q_2, q_3, \dots, q_N\} \quad (4)$$

4.3 Joint representation

The main task in VQA is to be able to combine the features of the two modalities by co-attending to each of them in order to bring out more meaningful representations from both the features. We use alternating co-attention defined in [24] for attending to both question and image together. Here, we alternate between question and image sequentially to generate image and question attention. This can be explained using following three steps.

- 1) First summary of Question is generated as a single vector q ;
- 2) Then using the question summary q , the image is attended;
- 3) Then using the attended image, the question is attended.

As shown in Fig. 6, an alternate attention function is defined as $\hat{x} = \text{Att}(X; g)$. The function takes two inputs: X and g . X is the Image features v (or a question features q) and g is the attention guidance that is derived from question (or image). The output of the function is the attended image features \hat{v} or attended question features \hat{q} depending upon what is the input to the attention function. When input is image features i.e. $X = v$, the output is attended image features \hat{v} and when input is question features i.e. $X = q$ then output is attended question features \hat{q} . The entire attention process can be explained using equation (5) – (7).

$$H = \tanh(W_x X + (W_g g)1^T) \quad (5)$$

$$at^x = \text{softmax}(w_{hx}^T \cdot H) \quad (6)$$

$$\hat{x} = \sum at_i^x x_i \quad (7)$$

Here 1 is a vector with all elements to be 1, at^x is the attention weight vector of feature X . $W_x, W_g \in R^{k \times d}$ and, $W_{hx} \in R^k$ are parameters. As shown in Fig. 6, we first give question features as an input i.e. $X = Q$ to the co-attention module, and attention guidance g is set to 0; In second step, we give image features as an input to the

attention module i.e. $X = V$ and the guidance g is an intermediate attended question feature \hat{s} from the first step; finally, the attended image feature \hat{v} is used as the guidance to attend the question again, i.e., $X = Q$ and $g = \hat{v}$.

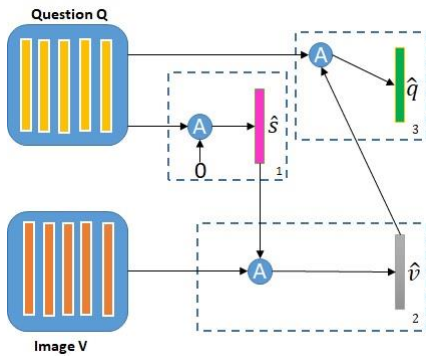


Fig. 6. Alternating co-attention mechanism [24]

4.4 Answer Generation

In most of the literature, the VQA problem is converted to the classification problem. This can be done by taking the most frequently occurring top 1000 answers in the VQA dataset as 1000 output classes [1]. Given an (image, question) pair, the model tries to find out the most probable answer out of these 1000 classes. In our implementation, instead of taking top 1000 candidate answers we have considered all those correct answers that appear more than 8 times in the training answer set, amounting to $N=3129$ candidate answers [28]. As a final step of VQA model, we concatenate output from all three attention modules and then fed to the MLP to generate final answer. The entire concatenation and the MLP is expressed using equation (8) – (11).

$$h^x = \tanh(W_x [\hat{q}^x + \hat{v}^x]) \quad (8)$$

$$h^y = \tanh(W_y [\hat{q}^y + \hat{v}^y], h^x) \quad (9)$$

$$h^z = \tanh(W_z [(\hat{q}^z + \hat{v}^z), h^y]) \quad (10)$$

$$A = \text{softmax}(W_h h^z) \quad (11)$$

Where \hat{q} and \hat{v} are attended question features and image features respectively and superscript x,y,z denote attended features from three levels respectively.

4.5 The Proposed Model Architecture

The proposed model shown in Fig. 7 is designed to take

advantage of BERT's ability to act as a powerful language feature extractor for sentences. First, image features are extracted using F-RCNN i.e. using Bottom-Up attention with adaptive K (10-100) as given in [28]. The dimensions of extracted image features are 100×2048 . We adopt a three-level hierarchical attention approach (see Fig. 6) as used in [24] between multiple co-attention modules. For the proposed model, Alternating Co-Attention method is used as it provides better performance to Parallel Co-Attention.

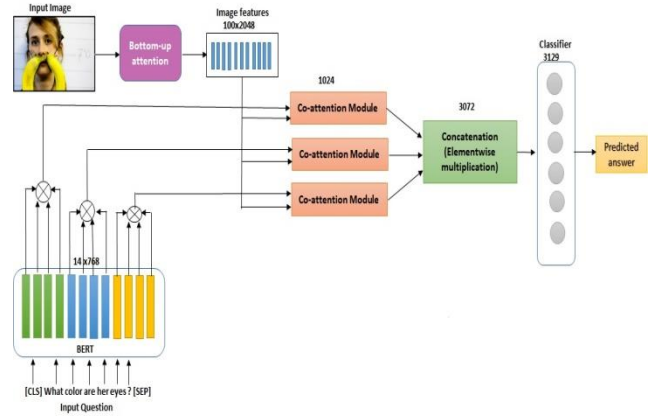


Fig. 7. Architecture of proposed BERT based Hierarchical Alternating Co-Attention VQA

Instead of taking the final output of BERT, features from 4 consecutive encoder layers of BERT are concatenated and passed as input features to each of three Co-Attention modules, attending to the same image features. This is based on the fact that each encoder layer of the transformer attends to features from the word to a phrase and ultimately to a sentence-level representation. The concatenated question features are fed to a Bi-LSTM whose output is fed to the alternating Co-Attention module to get the attended question and image features which are then combined using element-wise multiplication in order to obtain the joint question-image features. The output features are then concatenated and sent to a simple MLP for generating the output. Output was also inspired by the Top-Down Attention [28] where we take the top 3129 answers and use soft labels with BCE loss for training. We trained the model for 15 epochs. Initial learning rate greatly affects the model performance so, we set it to $2e-5$, following the learning rate schedule suggested in the MCAN [29]. The hidden size is set to 1024 and the dropout to 0.2 for the co-attention module. Fig 8. Shows the model summary.

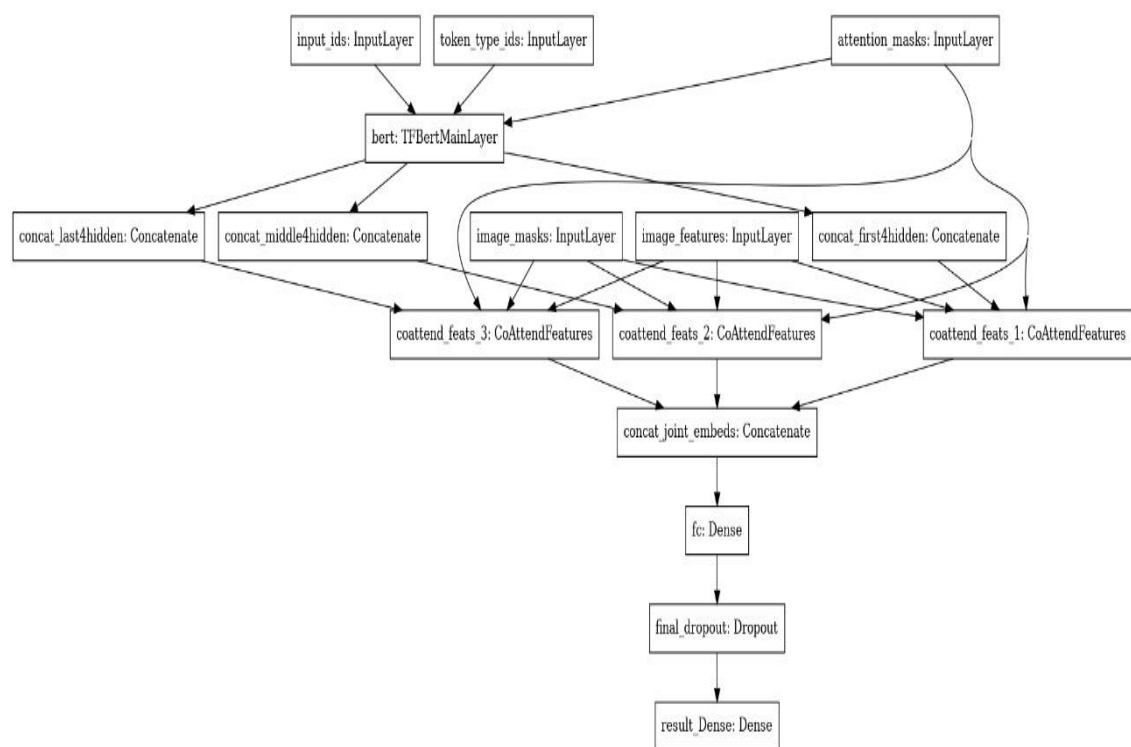


Fig. 8. Proposed Model Summary

Till now human accuracy on VQA dataset is ~81%. We obtain a validation accuracy of 63.94%, much higher than both Hierarchical Co-Attention [24] and the original Top-Down Bottom-Up attention [28] by 9.37% and 0.74% respectively, proving that the Co-Attention module is still effective when used with the improvements made to the training process as well as with the richer features obtained from the questions and images.

4.6 Data Set

We trained and tested our model on a benchmark dataset VQA2.0 that contains 204,721 images taken from the Microsoft COCO image set. It contains almost 1,105,904 questions in a natural language about these images. The dataset contains 3 questions per image and, 10 free-form human generated answers for each of these questions. The training set contains 443,757 questions on 82,783 images with 4,437,570 answers.

4.7 Data Pre-Processing

Developing a VQA models demands a lot of data pre-processing, since three different types of inputs are used for training a model: Image, Question and answer. Each image is identified by a unique image ID and corresponding each question has unique question ID. The image ids, related questions, and corresponding answers for each input were extracted and stored in separate JSON files. We select the answer that occurred the most among the 10 labelled answers to be the final answer. For testing the model on a test-set, we extract all 10 answers for each question and then model is evaluated using the evaluation

metric proposed in [1]. The BERT’s full tokenizer is used to tokenize the questions. As mentioned earlier, “[CLS]” and “[SEP]” tokens were added to each question sentence. We used pre-trained BERT model [20] with 12 layers and a hidden size of 768 for question embedding.

Algorithm:**Input:** (I,S) (Here I - Image and S - question sentence)**Output:** A (Generated answer)

1. Extract bottom-up image features by applying F-RCNN to the visual input image V. The output feature vector is of size $k \times 2048$, where k is the number of image locations.

$$v = f(I) = \{v_1, v_2, \dots, v_k\}, v_i \in R^k$$

Where v_i the feature of i^{th} object and, function is (*) indicates F-RCNN model. Here 'K' represents the number of object regions detected and 'k' represents the size of each object feature. Thus extracted image feature vector is represented as $v \in R^{K \times k}$, $k = 2048$ and $K=100$ (Variable between 10 to 100)

2. Extract question features using BERT.

Given a question sentence S containing N number of words, the question sentence S is represented as

$$S = \{s_1, s_2, s_3, \dots, s_N\}$$

Question Pre-processing:

- 2a. Tokenize Questions using BERT Tokenizers.
- 2b. Add [CLS] and [SEP] tokens to the sentence.
- 2c. Apply WordPiece tokenizer to split words into full forms or word pieces
- 2d. Pad each Question to a max length of 14.

The output of BERT is given as, $q = BERT(S) = \{q_1, q_2, q_3, \dots, q_N\}$

3. Concatenate output of 4 consecutive layers of BERT encoder to generate three-level of question feature vectors. (+ indicates concatenation)

$$q^{1-4} = (q_i + q_{i+1} + q_{i+2} + q_{i+3}), i = 1$$

$$q^{5-8} = (q_i + q_{i+1} + q_{i+2} + q_{i+3}), i = 5$$

$$q^{9-12} = (q_i + q_{i+1} + q_{i+2} + q_{i+3}), i = 9$$

4. Each feature vector q^{1-4} , q^{5-8} and q^{9-12} is then fed to the Bi-LSTM.

$$q^x = biLSTM(q^{1-4}), q^y = biLSTM(q^{5-8}), q^z = biLSTM(q^{9-12})$$

5. Output of Bi-LSTM and Image feature is then fed as an input to the alternating attention module taking image (or question) feature vector as X and attention guide g that is derived from question (or image) as input. The output is the attended image of question features, \hat{x} .

$$H = \tanh(W_x X + (W_g g)1^T)$$

$$a^x = \text{softmax}(w_{hx}^T \cdot H)$$

$$\hat{x} = \sum a_i^x x_i$$

Where 1 is a vector with all elements to be 1. $W_x, W_g \in R^{k \times d}$, $W_{hx} \in R^k$ are parameters. a^x is the attention weight of feature X. The alternating co-attention process is illustrated in Fig. 6. The output of this stage is attended image features \hat{v} and attended question features \hat{q} (refer Fig. 6)

6. Output of each co-attention module is combined using element-wise multiplication in order to obtain the joint question-image features.

$$h^x = \tanh(W_x [\hat{q}^x + \hat{v}^x])$$

$$h^y = \tanh(W_y [\hat{q}^y + \hat{v}^y], h^x)$$

$$h^z = \tanh(W_z [(z + \hat{v}^z), h^y])$$

7. The output is then fed to the simple Multi-layer Perceptron (MLP) which is a softmax classifier that calculates the probabilities of each answer from 3129 classes.

$$A = \text{softmax}(W_h h^z)$$

4.8 Model Parameters

We have trained the BERT based VQA model for 15 epochs and tuned hyper parameters to obtain the best results. Model parameters have been listed in Table 1.

Table 1: Proposed model parameters

Parameter	Value
Number of BERT layers	12
Hidden size	768
Dropout rate	0.2
Number of epochs	15
Optimizer	ADAM
Activation function	Relu
Batch size	64
Initial Learning rate	2e-5

5. Results and Discussion

The accuracy of the VQA model is computed as given in [1]. Accuracy is computed as,

$$Accuracy = \text{minimum}(\text{number of humans provided the model predicted answer}/3, 1)$$

(12)

This means the answer is correct if at least 3 humans said the predicted answer. Table 2 demonstrates the comparison of our experimental result evaluated on VQA 2.0 validation set with other baseline VQA models. We obtain a validation accuracy of 63.94%, much higher than both Hierarchical co-attention VQA model [24] and the original Top-Down Bottom-Up [28] by 9.37% and ~0.74% respectively, proving that the Co-Attention module is still effective when used with the improvements made to the training process as well as with the richer features obtained from the questions and images.

Table 2. Accuracy of different VQA Models on VQA v2.0 validation set.

Method	Yes/No	Number	Other	All
Hei-Co-Atten[24] reported in [29]	71.80	36.53	46.25	54.57
MCB [26] reported in [30]	77.37	36.66	51.23	59.14
Bottom-up Top-Down Attention [28]	80.07	42.87	55.81	63.20
BERT based Hie Alternate Coattention + Bottom up features (Ours)	81.88	43.02	55.84	63.94
Human Accuracy	95.49	80.84	67.89	80.78%

From Table 2 and Fig 9, it has been observed that the proposed VQA model improves upon original Hierarchical co-attention model [24] by 10.08%, 6.49% and 9.59% and by 1.81%, 0.15, 0.03% on [28] for Yes/no, number and other types of questions respectively.

It is clear from the results that the model needs to be improved for counting (i.e. number) type of question. Some of the sample outputs of our proposed model has been shown in Fig 10.

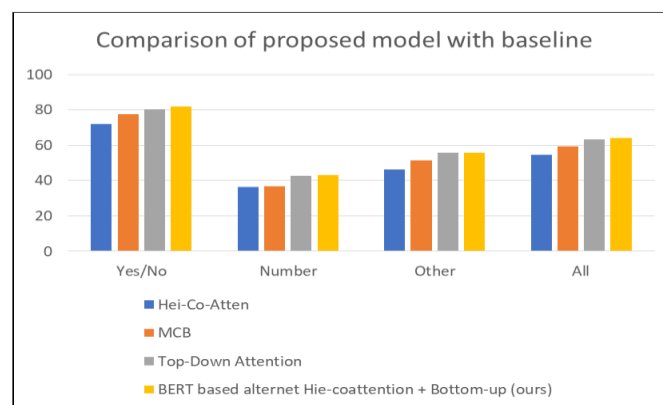


Fig. 9. Comparison of proposed method with baseline methods










 Image: 262162 Q: What colour is the bedspread? Answer (Predicted): beige ✓ Ground truth: beige	 Image: 383066 Q: Is her white container full of food? Answer (Predicted): no ✓ Ground truth: no	 Image: 120961 Q: Why is the book on the table? Answer (Predicted): reading ✓ Ground truth: reading
 Image: 121123 Q: How many people are on the elephant? Answer (Predicted): two ✓ Ground truth: two	 Image: 435384 Q: What kind of wine? Answer (Predicted): white ✓ Ground truth: white	 Image: 515555 Q: What time does the clock say? Answer (Predicted): 1 ✗ Ground truth: 1:35
 Image: 566543 Q: What game is being played here? Answer (Predicted): soccer ✓ Ground truth: soccer	 Image: 34556 Q: Do this bikes help our environment? Answer (Predicted): yes ✗ Ground truth: no	 Image: 435481 Q: Does the train have lights on? Answer (Predicted): yes ✓ Ground truth: yes

Fig. 10. Sample results of our proposed VQA model

Fig. 11 to 13 shows screen shots of Graphical user interface developed for the proposed VQA model. First step is to select any one image from the randomly displayed images. Second step is to enter desired question related to the image and in third step answer probability has been displayed. In the shown example, the probability of the answer being 2 is 44. 12%, 1 is 14.1% and so on...

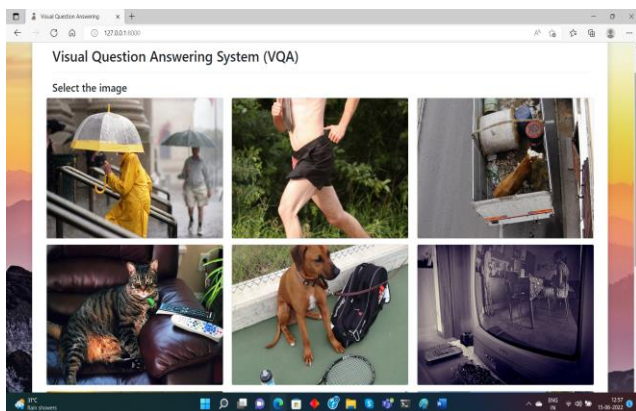


Fig 11: Selection of image

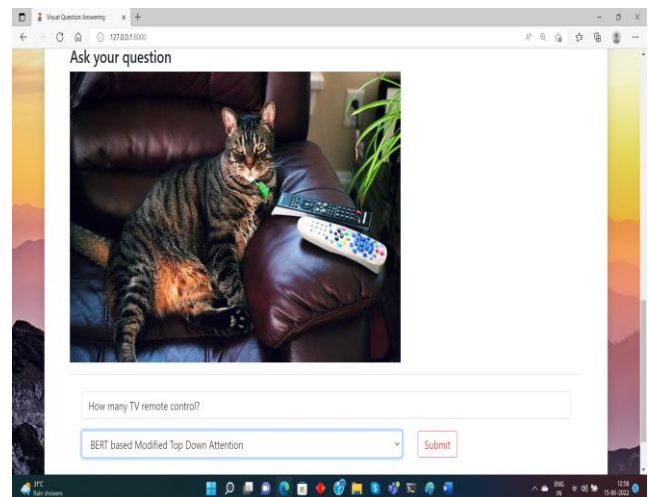


Fig. 12. Enter input question

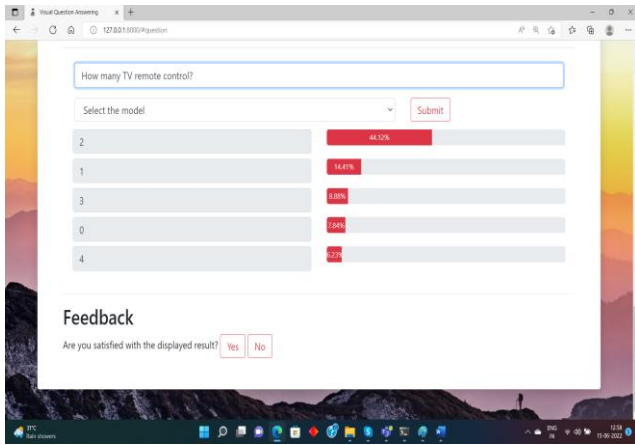


Fig. 13. Results – Answer probability

6. Conclusion

We proposed a novel BERT-based hierarchical VQA model using Bottom-up features and an alternate co-attention mechanism. We have explored the power of transformer BERT as a powerful language feature extractor for VQA. The proposed model improves the question feature extraction module using BERT and adopting a hierarchical approach for joint representation of image and question features. Our model takes the advantage of BERT's capability of extracting hierarchical features from a question by concatenating output of four consecutive BERT layers. We proved that bottom-up features with BERT based hierarchical approach outperforms both simple hierarchical VQA model and Top-down bottom approach for VQA.

References

- [1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C.L. Zitnick, D. Parikh, "VQA: visual question answering", Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2425–2433. doi: 10.1109/ICCV.2015.279.
- [2] N. Ruwa, Q. Mao, L. Wang and M. Dong, "Affective Visual Question Answering Network," IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2018, pp. 170-173, doi: 10.1109/MIPR.2018.00038.
- [3] Z. Yang, X. He, J. Gao, L. Deng and A. Smola, "Stacked Attention Networks for Image Question Answering," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 21-29, doi: 10.1109/CVPR.2016.10.
- [4] H. Noh, P. H. Seo and B. Han, "Image Question Answering Using Convolutional Neural Network with Dynamic Parameter Prediction," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 30-38, doi: 10.1109/CVPR.2016.11.
- [5] Ilija Ilievski, Shuicheng Yan, Jiashi Feng, "A Focused

Dynamic Attention model for visual question answering." [Online]. Available: <https://arxiv.org/abs/1604.01485> (2016).

- [6] K. Kafle and C. Kanan, "Answer-Type Prediction for Visual Question Answering," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4976-4984, doi: 10.1109/CVPR.2016.538.
- [7] 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).
- [8] M. Malinowski, M. Rohrbach and M. Fritz, "Ask Your Neurons: A Neural-Based Approach to Answering Questions about Images," IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1-9, doi: 10.1109/ICCV.2015.9.
- [9] Haoyuan gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu, "Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering", Proc. Advances in Neural Inf. Process. Syst., 2015.
- [10] Bolei Zhou, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus, "Simple baseline for visual question answering", [online] Available: arXiv preprint arXiv:1512.02167, 2015.
- [11] Xu, H., Saenko, K., "Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering" Computer Vision – ECCV 2016. Lecture Notes in Computer Science, vol. 9911. Springer, Cham. https://doi.org/10.1007/978-3-319-46478-7_28
- [12] Gupta D., Lenka P., Ekbal Asif, Bhattacharya P., "A Unified Framework for Multilingual and Code-Mixed Visual Question Answering", Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (pp. 900–913). Association for Computational Linguistics (2020).
- [13] 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).
- [14] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In ICCV, pages 1839–1848, 2017
- [15] Pan Lu, Hongsheng Li, Wei Zhang, Jianyong Wang, and Xiaogang Wang. "Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering", In Proceedings of the AAAI Conference on Artificial Intelligence, 2018, pp 7218-7225.

- [16] Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel, "Explicit knowledge-based reasoning for visual question answering", *IJCAI*, pages 1290–1296, 2017.
- [17] Jin-Hwa Kim, Jaehyun Jun, & Byoung-Tak Zhang, "Bilinear Attention Networks", 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada. CoRR, abs/1805.07932.
- [18] Lianli Gao, Liangfu Cao, Xing Xu, Jie Shao, Jingkuan Song, "Question-Led object attention for visual question answering", *Neurocomputing*, Volume 391, 2020, Pages 227-233, ISSN 0925-2312
- [19] Deepak Gupta, Pabitra Lenka, Asif Ekbal, and Pushpak Bhattacharyya, "A Unified Framework for Multilingual and Code-Mixed Visual Question Answering", In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 900–913, Suzhou, China. Association for Computational Linguistics.
- [20] Attention Is All You Need, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
- [21] K. J. Shih, S. Singh and D. Hoiem, "Where to Look: Focus Regions for Visual Question Answering," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4613-4621, doi: 10.1109/CVPR.2016.499.
- [22] R. Burt, M. Cudic and J. C. Principe, "Fusing attention with visual question answering," 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 949-953, doi: 10.1109/IJCNN.2017.7965954.
- [23] Nguyen, Kien & Okatani, Takayuki. (2018). Improved Fusion of Visual and Language Representations by Dense Symmetric Co-attention for Visual Question Answering. 6087-6096. 10.1109/CVPR.2018.00637.
- [24] 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).
- [25] G. Gu, S. T. Kim and Y. M. Ro, "Adaptive attention fusion network for visual question answering," *IEEE International Conference on Multimedia and Expo (ICME)*, 2017, pp. 997-1002, doi: 10.1109/ICME.2017.8019540.
- [26] 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", *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, Austin, Texas. Association for Computational Linguistics.
- [27] M. Dias, H. Aloj, N. Ninan and D. Koshti, "BERT based Multiple Parallel Co-attention Model for Visual Question Answering," 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), 2022, pp. 1531-1537, doi: 10.1109/ICICCS53718.2022.9788253.
- [28] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, "Bottom-up and top-down attention for image captioning and VQA", arXiv: 1707. 07998 (2017).
- [29] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, Qi Tian, "Deep Modular Co-Attention Networks for Visual Question Answering" *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6281-6290
- [30] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, "Making the V in VQA matter :Elevating the role of image understanding in visual question answer- ing", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6325–6334 .