Proceedings of the Sixth International Conference on Intelligent Computing and Control Systems (ICICCS 2022)
IEEE Xplore Part Number: CFP22K74-ART; ISBN: 978-1-6654-1035-9

2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS) | 978-1-6654-1035-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICICCS53718.2022.9788253

BERT based Multiple Parallel Co-attention Model for Visual Question Answering

1st Mario Dias
Dept. of Computer Engineering
Fr. Conceicao Rodrigues College of Engineering
Mumbai, India
mario.dias3100@gmail.com

2nd Hansie Aloj
Dept. of Computer Engineering
Fr. Conceicao Rodrigues College of Engineering
Mumbai, India
hansiealoj3@gmail.com

3rd Nijo Ninan
Dept. of Computer Engineering
Fr. Conceicao Rodrigues College of Engineering
Mumbai, India
nijoninan23@gmail.com

4th Dipali Koshti,
Dept. of Computer Engineering
Fr. Conceicao Rodrigues College of Engineering
Mumbai, India
dipalis@fragnel.edu.in

***Abstract* — Humans can easily interpret visual and textual content whereas for a machine this is a challenging task. Visual question answering is a well-known problem in the field of computer vision and NLP where an image and a question related to the image are given and the machine has to generate a natural language answer. This paper explores the use of transformer BERT as a language model in VQA. The BERT-based multiple parallel co-attention visual question answering model has been proposed and the effect of introducing a powerful feature extractor like BERT for language modeling has been studied. From the experimental results, it is concluded that the proposed model improves over the original baseline VQA model by 3%.**

Keywords— Visual Question Answering, Image Question answering, Hierarchical VQA, BERT

I. INTRODUCTION

Visual Question answering (VQA) is considered to be a multi-modal task as it requires knowledge from multiple fields. It is obvious that VQA requires expertise in multiple fields like Natural language processing, image processing, and knowledge reasoning. The problem is to generate a natural language answer given an (image, question) pair. Fig.1 shows an example of VQA [1].

Although for humans, it is easy to interpret the visual content present in an image and relate it to the textual content of a given question, it's a challenging task for a machine. The machine needs to understand the different modalities and relate them to generate correct answers. In general, there are four stages of any VQA model: extraction of question features, extraction of image features, a fusion of question and image features, and finally answer generation. Image feature extraction may require understanding the image at different levels. This may include a number of computer vision tasks such as object counting, object identification, scene detection, etc. Also, question feature extraction involves the extraction of important keywords from the question and relating them to a given image. The third phase, fusion requires the processing of question features and image features in a common space to explore the relationship between the two. Finally, phase 4 generating an answer depends on the mode of the answer to be generated. To simplify the answer generation, many authors have treated VQA as the classification problem [1],[2],[3],[4] by selecting 1000 (or 3000) most frequently occurring answers in a dataset as final classes. The VQA model generates the top 5 best matches among these 1000 (or 3000) classes.

This paper proposes two models that exploit the use of BERT (Bidirectional Encoder Representations from Transformers) in the VQA model. Our best model BERT + Multiple parallel Co-attention achieves 57.04% accuracy and improves on the baseline model Hierarchical Co-attention VQA [9].

Ques: How many slices of pizza are there?
Answer: 6

Ques: Where is this animal found?
Answer: Zoo

Fig. 1. Example of Visual question answering

II. RELATED WORK

A variety of models and solutions were proposed by different researchers to solve the VQA problem. The easiest solution was proposed by [1] to use CNN for image featurization, LSTM for question featurization, and simple concatenation or element-wise multiplication for a fusion. The VQA model in [1] considers image and question as a whole and failed to pay any special attention to important keywords of the question or important image regions related to a given question. VQA performance can be improved by incorporating an attention mechanism [5]. In [2] authors proposed a novel attention-based VQA model wherein they performed multiple-step reasoning on the image. Each attention network tries to locate the important visual information that leads to the correct answer. They proved that multiple-step reasoning improves the model's answer prediction capability. In [4] Ilievski et al. proposed a novel question-driven FDA model for VQA. Based on the

important keywords in a given question, important image regions were extracted and these region features were combined with the global image features using LSTM. These combined image features were then fused with question features. The question is encoded using the LSTM network. In [6] authors argued that knowing question type helps reduce the candidate answer space. The model employed Faster-RCNN to extract high-level image features and finds top k candidate objects. Question features were extracted using Bi-LSTM. The co-attention mechanism uses self-attended question features to attend to image features. In [7] Gao Lianli et.al used both global and local features of the image. Global features were obtained using ResNet and local (object level) features were obtained using Faster RCNN. Important keywords of questions were employed to find important image objects. Only those objects that are related to questions were extracted and then combined with question features. Kien and Okatani [8] proposed a fully symmetric co-attention-based fusion model where each word in a question attends to an image and in turn, each image region attends to the question word. In [9] the authors extracted the word level, phrase level, and sentence level question features and fused each of them with the image features using two different co-attention mechanisms: Alternating co-attention and parallel co-attention. We use the hierarchical co- attention model [9] as our baseline and propose a few modifications in the model to improve model performance along with the ablation study.

III. PROPOSED MODEL

Given a visual image ‘v’ and a related question ‘q’ in natural language, the task of the VQA model is to generate the correct answer ‘a’ in natural language. The proposed system aims to design a complete visual question answering framework capable of reasoning open-ended questions and generating answer in natural language. We propose an improved BERT-based VQA model with the multiple parallel co-attention mechanism. Recently, the attention mechanism has changed the way we process natural language and revolutionized the deep learning models. The attention mechanism is not only used for natural language processing but also in fields like computer vision. In the context of VQA, the model needs to pay attention to important keywords in a question, and then these keywords may be used to attend to important parts (or objects) of the image.

We adapt the Hierarchical Co-Attention [9] as the base model. As shown in Fig. 3, the model proposed in [9] uses both the image and question attention mechanism to predict an answer. The model generates the three-level hierarchy (namely word, phrase, and sentence level) of the given question. For the word-level features, authors have used one-hot encoding to generate a word vector; for the phrase-level features, they applied 1D convolutional on the word embedding vector. And for the sentence level embedding, LSTM was used. Here authors have proposed two co-attention models - alternating co-attention and parallel co-attention. For our experiment, we select the model which uses **VGGNet** for image feature extraction and uses a **parallel co-attention** mechanism as shown in Fig. 2 for co-attending to the different features in our experimentation.

Transformers have shown promising results in the field of NLP. We explore the BERT-based transformers for language modeling in the VQA model due to its self-attention abilities. We proposed three BERT-based models keeping the base model architecture similar to Hierarchical Co-Attention [9]. In the end, we propose a newer modified BERT-based Multiple parallel co-attention VQA model, highlighting the effectiveness of its co-attention module in providing competitive performance alongside similar SOTAs.

We describe the major components of our base model in the following section:

(a) Image Feature Extraction

The image feature extraction model uses a CNN model to extract global image features from the given image. Based on the image model used by [2]. We use models trained on ImageNet like VGGNet and ResNet to extract the global feature vectors from the last pooling layer so that we can retain the spatial information. In order to obtain more meaningful features from the images, we resize them to 448x448 and then take the features from the last pooling layer. The dimensions of the image features extracted would be 512x14x14 for VGGNet which we compress to generate 512x196 features per image. For the model that uses ResNet features, the dimensions are 2048x14x14 instead.

(b) Question Feature Extraction

To extract the semantic meaning of the text from the questions, we propose to utilize and adapt the process of hierarchical embeddings at the word, phrase and, question level as used in [9]. For this, the questions are first tokenized using NLTK tokenizers to extract the words which are then embedded into vector space using Word2Vec embeddings. Another proposed method is to utilize pre-trained BERT (Bidirectional Encoder Representations from Transformers) models [10] to generate the word-level feature vectors. The features vectors produced by BERT would take the entire sentence into account to extract semantic meaning for the given word and thus will provide more meaningful features to extract from. These word-level features are then taken and 1-D convolution is applied to them as proposed in [9]. At each word location, we compute the inner product of the word vectors with filters of three window sizes: unigram, bigram, and trigram. After the convolutional result, we then apply max-pooling across different n-grams at each word location to obtain the phrase-level features. To obtain the question level features, we use an LSTM to encode the phrase level features obtained. The corresponding question-level features are the LSTM hidden vectors.

(c) *Joint Feature Attention Model*

For combining the visual and textual features we make use of attention map mechanisms based on the co-attention mechanisms discussed in [9]. We create three attention maps that attend to the image and each question embedding level separately and then join them together using a final joint attention map. By attending to each question embedding level separately, the model will learn to focus and pay attention at different levels in both the image and the question thus enhancing the features of interest when we combine the attention maps. We currently propose to adapt the parallel co-attention mechanism described in [9] which co-attends to both the question and the image simultaneously. This is done at each level in the question hierarchy. Fig. 2 gives us the representation of the co-attention mechanism. Here $\boldsymbol{V}$ denotes the image features and $\boldsymbol{Q}$ the question. $\boldsymbol{C}$ denotes the affinity matrix which is calculated by

$$\boldsymbol{C} = \boldsymbol{tanh}(\boldsymbol{Q}^{\boldsymbol{T}}\boldsymbol{W}_{\boldsymbol{b}}\boldsymbol{V})$$

Where Wb contains the weights. Once calculated, this affinity matrix is then considered as an additional feature and used to predict the image and attentions maps via the following

$$\boldsymbol{H}_{\boldsymbol{v}} = \boldsymbol{t}(\boldsymbol{W}_{\boldsymbol{v}}\boldsymbol{V} + (\boldsymbol{W}_{\boldsymbol{q}}\boldsymbol{Q})\boldsymbol{C}) \qquad \boldsymbol{av} = \boldsymbol{softmax}(\boldsymbol{w}^{\mathrm{T}}{}_{\boldsymbol{hv}}\ \boldsymbol{H}^{\boldsymbol{v}})$$
$$\boldsymbol{H}_{\boldsymbol{q}} = \boldsymbol{tanh}(\boldsymbol{W}_{\boldsymbol{q}}\boldsymbol{Q} + (\boldsymbol{W}\boldsymbol{v}\boldsymbol{V})\boldsymbol{C}^{\boldsymbol{T}}) \qquad \boldsymbol{aq} = \boldsymbol{softmax}(\boldsymbol{w}^{\boldsymbol{T}}{}_{\boldsymbol{hq}}\,\boldsymbol{H}^{\boldsymbol{q}})$$

where $\boldsymbol{W}_{\boldsymbol{v}}$, $\boldsymbol{W}_{\boldsymbol{q}}$, $\boldsymbol{w}_{\boldsymbol{hv}}$,$\boldsymbol{w}_{\boldsymbol{hq}}$ are the weight parameters, and $\boldsymbol{a}_{\boldsymbol{v}}$ and $\boldsymbol{a}_{\boldsymbol{q}}$ are the attention probabilities of each image region and word respectively

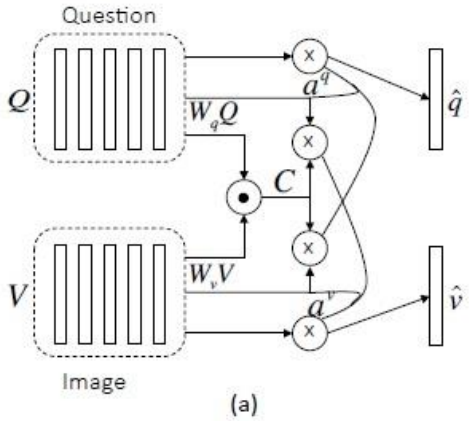


Fig 2: Parallel co-attention

d) *Answer Prediction Model:*

We treat VQA as a classification task, classifying it against the top 1000 most frequently occurring answers in the VQA v2.0 dataset. We predict the answers based on the co-attended image and question features which were combined using a joint attention map and feed it as input to a multi-layer perception (MLP) which outputs a vector of probabilities that will tell us the most probably answer from the possible 1000 answers. Based on the above base model, we implemented three different VQA models.

- *BERT + Hierarchical Co-Attention Model*
- *BERT+Hierarchical Co-Attention with Single Co-Attention*
- *BERT + Multiple parallel Co-Attention*

A. BERT + Hierarchical Co-Attention Model

We modified the original Hierarchical Co-Attention model by replacing the language model with BERT [10]. Here we adopt the same model as Hierarchical Co-Attention [9] (see Fig. 4.), extracting the word-level features using a pre-trained BERT model [12]. Note that here we have introduced BERT for the language model. We also adjust the dropout distribution and reduce the number of LSTMs used from 2 to 1 for sentence-level feature extraction to accommodate the superior contextual embeddings obtained from BERT. We then adjust the MLP (Multilayer perceptron) used in Hierarchical Co-Attention by adding another DenseNet to extract better features. For image feature extraction, the first input image is resized to 448x448, and then pre-trained VGGNet is used to extract image features.

The model uses parallel co-attention by attending to all the three features: sentence level, phrase level, and word-level feature parallelly as in the original Hierarchical Co-Attention [9].

B. BERT+Hierarchical Co-Attention with Single Co-Attention

This model shown in Fig. 5 is the same as the above model however, we adjust the dropout rates in this model to 0.3 from 0.5 and most remarkably utilize a single shared co-attention mechanism between each of the three hierarchical levels. This means the weights are shared between each level. Our intuition is based on the fact that BERT's contextual features provide much more meaningful insight compared to the separate co-attentions performed in Hierarchical Co-Attention.

C. BERT + Multiple parallel Co-Attention

We design a modified version of the original Hierarchical Co-Attention [9] model shown in Fig. 6 based on our findings from the previous models and conclude that much information is lost due to the phrase feature extraction method and LSTM methods utilized to extract more detailed information. Instead inspired by multi-head attention we utilize 3 identical parallel co-attention heads to extract features from BERT. We modify the Co-Attention module to include Layer normalization and use the ReLU activation function instead of Tanh. Another important step was to perform finetuning on BERT itself with the question input to improve the alignment between the BERT Embeddings and the result. We also replace VGGNet with ResNet-152 to extract image features as using ResNet proved to improve the accuracy of the VQA model [4],[9],[11]. We use a decreased learning rate of 3e-5 with Adam optimizer to ensure we don't lose information during the finetuning process. We train the model for 10 epochs with

57.4% accuracy on the validation set which shows how significant of a boost finetuning BERT helps even with standard NLP models. We used a hidden size of 512 and classified it among the top 1000 answers.

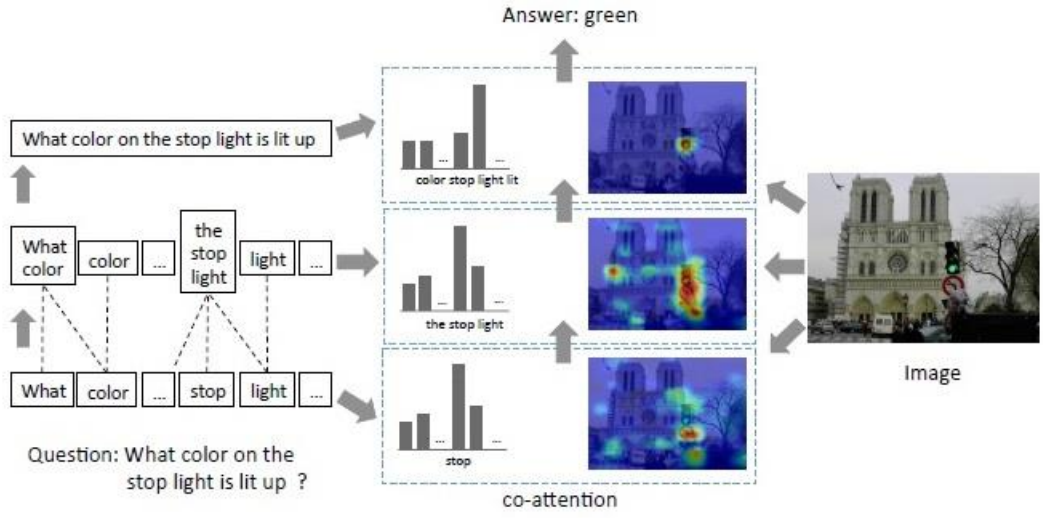


Fig. 3. Hierarchical Co-Attention Model [9]

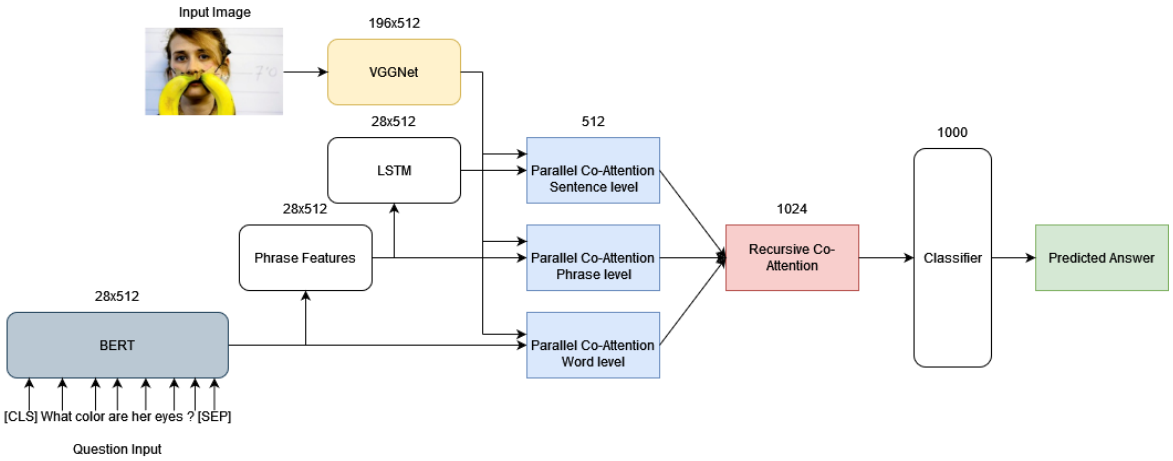


Fig. 4. Model architecture for BERT + Hierarchical Co-Attention model

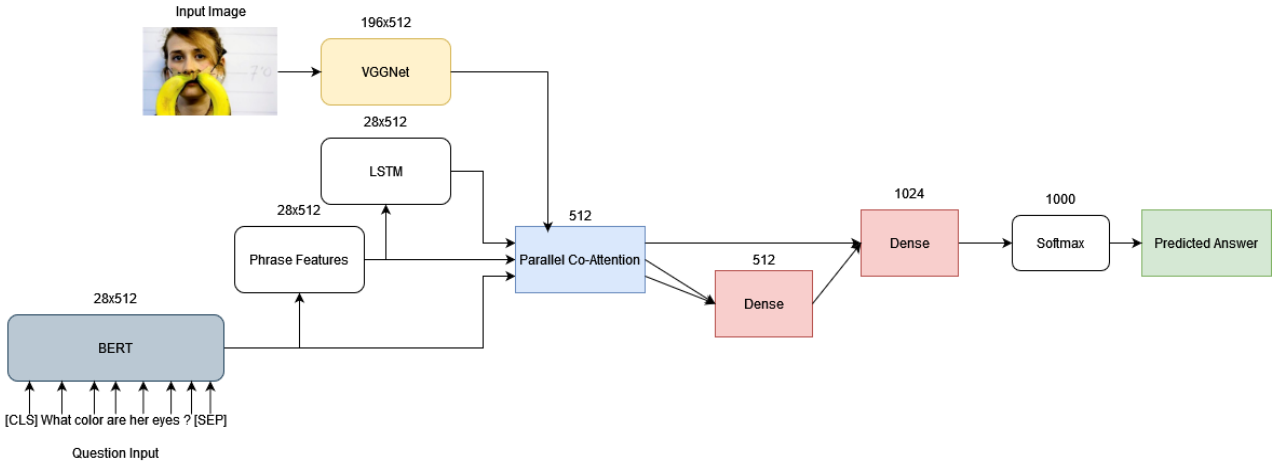


Fig. 5. Model architecture for BERT + Hierarchical Co-Attention with Single Co-Attention model.

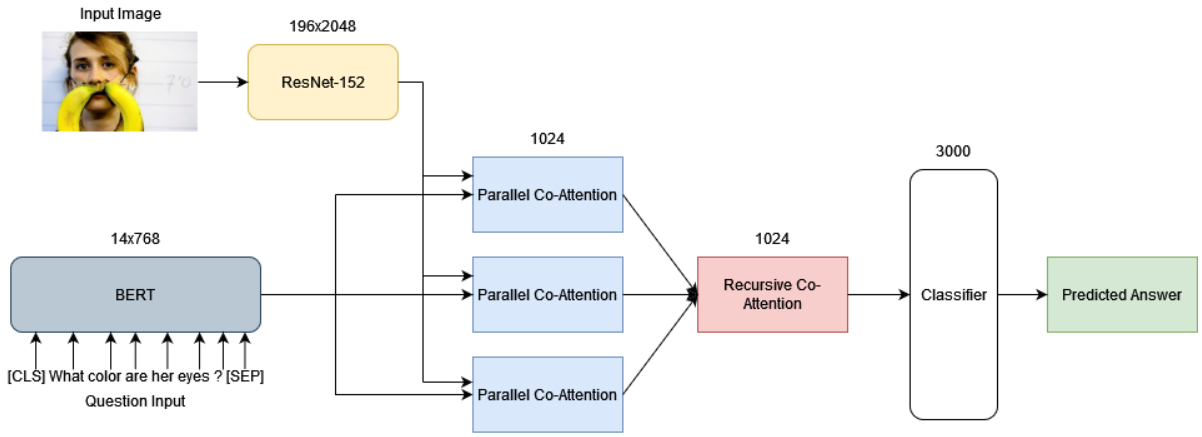


Fig. 6. Model architecture for BERT + Multiple Parallel Co-Attention model

D. Data Set:

The proposed model is trained and evaluated on the VQA2.0 dataset [13]. The dataset contains over 204,721 images coming from COCO images. It contains 614,613 natural language questions where each image contains 3 questions and each question contains 10 free-form but concise answers labeled by human annotators. The VQA dataset was revised in 2017 [13] to counter the original dataset's strong language bias. VQA v2.0 consists of 443,757 training questions on 82,783 images taken from the COCO dataset which consists of a total of 4,437,570 answers.

E. Data Preprocessing

We extract the questions, image ids, and answers for each input and store them separately. For the answers, we select the modal answer as the answer which occurred the most among the 10 labeled answers to be the official answer. For the testing set, we extract all 10 answers for each question in order to evaluate the models on the evaluation metric proposed in [1]. As mentioned in other literature, we have also converted the VQA problem into a classification problem by selecting the 1000 most frequently occurring answers in a dataset as final classes. Our VQA model generates the answer by selecting the most suitable class among these 1000 classes for the given question, image pair. After selecting the data corresponding to the 1000 most frequent answers, we tokenize the questions using NLTK libraries for baseline models or using the Full Tokenizer provided by BERT in models which used it. BERT models required the input questions to contain "[CLS]" and "[SEP]" tokens, which increased the overall size of each question. Most of the questions in a dataset consisted of less than 10 words, with the most being 22. Hence, for our models, the maximum padding length for each question was between 26 and 30, depending on the model being used. Answers were encoded using a pretrained label encoder to ensure consistency among the different models. Images are first resized to 448x448 and then image features are extracted using pre-trained VGG19. To obtain the image features of size 512x196, the features from the last pooling layer of VGG19 net have been extracted. For

BERT-based hierarchical VQA models, BERT is used to extract the language (question) features. The pre-trained BERT model from [12] with 12 layers and a hidden size of 512 to was used to better work with our implementation. For models which use GloVe vectors, we select the GloVe vectors with a feature size of 300 trained on the Common Crawl corpus.

IV. Results and Discussion

We use accuracy as defined in [1] by Antol S. to evaluate and compare our model with the baseline. For VQA, computing accuracy is not straightforward since the answer generated by the model and the ground truth answer may not be exactly the same but still correct. For example, the answer "tree" generated by the VQA model when compared with the ground truth answer "oak tree", does not mean that the generated answer is absolutely incorrect. For example, the accuracy is computed as mentioned in [1] as,

Accuracy = min (# humans that provided that answer/3,1)

So, as mentioned in the above equation an answer is considered accurate if at least 3 human workers provided that exact answer.

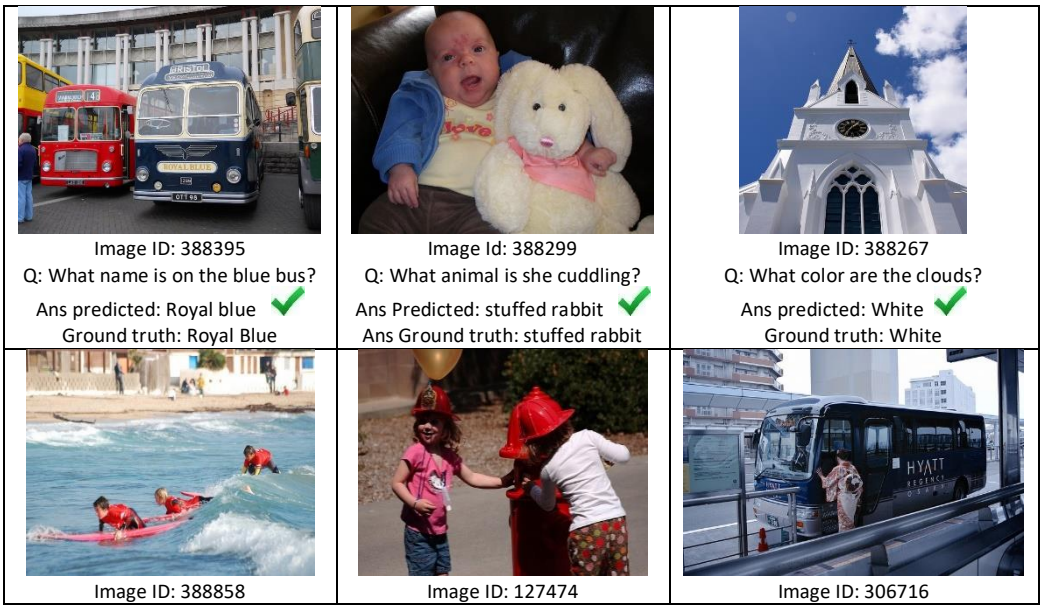
TABLE I: EVALUATION OF VQA MODELS ON VQA 2.0

Method	Yes/No	Number	Other	All
Hierarchical Co-Attention [9] (Baseline)	71.80	36.53	46.25	54.57
BERT + Hierarchical Co-Attention	69.36	34.61	44.4	52.49
BERT + Hie-Co-Attention with Single Co-Attention	73.26	36.79	43.66	54.03
BERT + Multiple Co-Attention	**76.44**	**37.24**	**48.15**	**57.84**

Table I shows the result of the proposed three models on the validation set of the VQA 2.0 dataset. Our experimental results show that BERT + Multiple Co-attention models improve over the original Hierarchical Co-Attention [9] by 3%. Out model shows improvement of 4.64%, 0.71%, and 1.9% for Yes/No type of question, Number (i.e. counting) questions, and other types of questions respectively. Fig. 7 shows a few of the sample images, related questions, and predicted answers by our best model.

V. CONCLUSION

In this paper we observe the effect of introducing BERT as a question feature extractor in the Hierarchical Co-Attention model, observing different ways in which performance could be improved on the VQA task. We modify our approach and design based on improvements suggested in more recent papers, especially in the training process as well as utilizing better image features and feature combining methods in order to better represent the joint features of the image and question. Thus, our final model BERT + Multiple Co-Attention is able to achieve competitive results to the SOTA models we used as baselines and showcased how robust the co-attention module utilized in Hierarchical Co-Attention is when used alongside powerful feature extractors for questions like BERT. The proposed model can be further improved by developing a complete BERT-based VQA model using BERT as a visual feature extractor.



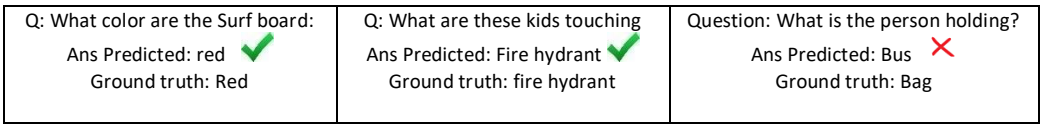


Fig 7. Sample predicted answers and ground truth on VQA 2.0 validation images.

REFERENCES

[1] S. Antol et al., "VQA: Visual Question Answering," IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2425-2433, doi: 10.1109/ICCV.2015.279.

[2] SAN : Z. Yang, X. He, J. Gao, L. Deng and A. Smola, "Stacked Attention Networks for Image Question Answering," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 21-29, doi: 10.1109/CVPR.2016.10.

[3] DPP: H. Noh, P. H. Seo and B. Han, "Image Question Answering Using Convolutional Neural Network with Dynamic Parameter Prediction," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 30-38, doi: 10.1109/CVPR.2016.11.

[4] FDA: Ilija Ilievski, Shuicheng Yan, Jiashi Feng,"A Focused Dynamic Attention model for visual question answering." [Online]. Available: https://arxiv.org/abs/-1604.01485 (2016)

[5] Bahdanau, Dzmitry, Kyunghyun Cho and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate." CoRR abs/1409.0473 (2015): n. pag..

[6] Chao Yang , Mengqi jianbin Jiang , Weixin Zhou and Keqin li ."Co-Attention Network With Question Type for Visual Question Answering," in IEEE Access, vol. 7, pp. 40771-40781,201 (2019).

[7] Gao Lianli, Cao Liangfu, Xu Xing, Shao Jie, Song Jingkuan, "Question-Led object attention for visual question answering," Neurocomputing, Volume 391, 2020, Pages 227-33, ISSN 0925-2312 (2020).

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

[9] Hei-co-atten: 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).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova," BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

[12] Pretrained BERT: Turc, I., Chang, M.W., Lee, K., & Toutanova, K. (2019). Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. arXiv preprint arXiv:1908.08962v2

[13] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the role of image understanding in visual question answering. In CVPR, pages 6325–6334. IEEE Computer Society, 2017.