Proceedings of the Third International Conference on Artificial Intelligence and Smart Energy (ICAIS 2023)
IEEE Xplore Part Number: CFP23OAB-ART; ISBN: 978-1-6654-6216-7

Knowledge Blended Open Domain Visual Question Answering using Transformer

1st Dipali Koshti
Department of Computer Science and Engineering
Sir Padampat Singhania University
Udaipur, Rajasthan, India
dipali.koshti@spsu.ac.in

2nd Dr. Ashutosh Gupta
Department of Computer Science and Engineering
Sir Padampat Singhania University
Udaipur, Rajasthan, India
ashu.gupta@spsu.ac.in

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

2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS) | 978-1-6654-6216-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAIS56108.2023.10073911

***Abstract*—Interacting with an image in the form of dialog is one of the challenging applications of the vision-language model. Image question answering allows us to interact with an image in form of question and answer. Ask any question about the image and the machine will generate an answer in a natural language. Not all questions are image-dependent; some of the questions may require external knowledge. Integrating external knowledge in an image question-answering model has been an open research area. A novel knowledge-incorporated image question-answering model based on a transformer using deep co-attention has been proposed. The model leverages the structured knowledge present in the ConceptNet. Important objects from the image and important keywords from the question are extracted. Using these extracted objects and text keywords, related concepts from the ConceptNet have been extracted. The top five most related concepts have been considered for further processing. A novel attention mechanism using a transformer has been introduced to combine this external knowledge with the Visual question answering model. The proposed model is evaluated based on VQA 2.0 dataset. The experimental results show that the incorporation of the external knowledge base in the VQA model allows the model to answer more complex open-domain questions and achieves the accuracy of 67.97% on VQA validation set.**

Keywords—Image question answering, Bidirectional Encoder Representation from Transformers, Visual question answering, ConceptNet, Knowledge visual question answering

I. INTRODUCTION

Developing a model for Visual question answering (also called Image question answering) introduced by [1] has remained a challenge as it requires the incorporation of human intelligence into the machine. The task of the VQA model is to generate a natural language answer for a question posed for a given image. Humans can perform a high level of reasoning over visual and textual data and can relate to the context present in the question. Humans can also apply common sense knowledge learned from past experiences to answer the question. With the advancement in Deep learning and the natural language processing field, it is now possible to perform human-like reasoning on multiple modalities.

In the beginning, most of the researchers used CNN and pre-trained networks like VGGNet, ResNet for image feature extraction [2]-[4] and RNN and its variations [5]-[7] for question feature extraction and combined these features to perform joint comprehension. Once the image features and question features are extracted, the next step is to combine (or fuse) these two features. Different methods such as simple concatenation to elementwise multiplication [1], [6], [8] have been used for fusion.

After the introduction of the attention mechanism by [9], most of the VQA models introduced various attention mechanisms to solve the VQA problem. The attention mechanism allows one modality to attend to the other modality. In the case of VQA, a given question can be used to attend to the important image regions. In other words, only those parts of the image will be attended to that are important for a given question, the rest are ignored. Alternatively, a co-attention mechanism can be used, wherein each question word attends to the important parts of the image and each image region attends to the important words of the question [10], [11]. Although the attention mechanism drastically improved the VQA performance, these models failed to answer the open-domain questions that require external knowledge.

To address this problem, this study proposes a novel knowledge-incorporated image question-answering model capable of reasoning open-domain questions. The model leverages the vast common sense knowledge available in ConceptNet and integrates it into the proposed VQA framework. The image features and object labels are extracted by using F-RCNN. The latest transformer technology (BERT) has been explored in order to extract question features and knowledge features. It also uses the transformer's decoder strategy for performing joint comprehension of image and fact features together with the question features. The top 5 object labels of the image and the top 3 object-related keywords of questions have been used in order to extract related knowledge from the ConcepNet. The experimental results prove that incorporating external knowledge improves VQA performance as it allows us to answer the knowledge-based question also.

II. RELATED WORK

Visual question answering first introduced by [1], is a good attempt to combine computer vision and natural language processing as it requires processing and comprehending both image content and language content. Most of the VQA models use only two modalities: an image and a question in order to generate an answer. They focus only on visual aspects present in the image to generate the answer. But in practice, not all the questions can be answered based on only visual aspects; they need information beyond the content present in the image and question.

Many authors used a Graph-based methods for extracting external knowledge from the knowledge base. Also, the use of a variety of knowledge bases such as WikiData, WikiPedia, DBPedia, and conceptNet can be observed in the previous work. In [12] Guohao et al. leveraged the external knowledge by extracting the first-

order sub-graph from ConceptNet by extracting those nodes in the conceptNet that match the important entities in an image and a question. This subgraph is further refined by ranking the edges according to their relevance to the question. Finally, the image and extracted knowledge features are stored in the dynamic memory module, and performing the joint understanding of these features using the attention mechanism generates the final answer. In [13] authors proposed a new VQA dataset called HVQR that conations an image, a pair of question, and answer, and the knowledge in the form of triplet. They claimed to solve the problem of multi-step reasoning. Some questions require multi-step reasoning to arrive at the answer. The model first converts the question into a structured query – a triple tree. The image features are parsed using F-RCNN. This query layout is then used to infer the common sense knowledge base and visual content in the image. In [14] Wu et al. used three modalities: Image, Question, and Image captions. The first textual representation of the image is generated using the caption generation method. Then top 5 objects from an image using CNN were extracted. SPARQL was used to query the DBPedia and relevant external knowledge was extracted. This external knowledge along with the visual and caption features are combined and fed to the LSTM which in turn generates an answer. [15] Proposed a graph-based VQA model that combines scene graph with concept graph. The authors used Sentence BERT to generate question-image representation and find the top K knowledge instances by using cosine similarity between knowledge and question-image representation. In [16] the model generates the initial scene graph from an image and the corresponding Knowledge graph from external knowledge bases. Based on the node matching in both graphs, candidate nodes from the knowledge graph are combined with the scene graph to obtain an enriched scene graph. All of the above methods used graphs; graph-based methods are more complex and slow to converge. Marino et al. [17] proposed a VQA model called ArticleNet that extracts related articles from WikiPedia and finds whether the answer appears in any article. While the method is able to successfully incorporate external knowledge, the method is expensive since WikiPedia is unstructured knowledge base. Retrieving relevant knowledge requires hand-crafted process such as collecting all possible articles, ranking those articles, filtering articles etc.

III. METHODOLOGY

A novel open-domain Visual Question Answering framework that integrates additional external knowledge from the publicly available knowledge bases to answer open-domain complex questions has been proposed. A detailed architecture of the proposed model has been shown in Fig. 1. The model contains the following five components: A Feature extraction module, Joint representation Module, Attention reduction module, Adaptive score generation module and a classifier.

A. Feature Extraction Module

This module contains three sub modules: One for extracting Image features, Second for extracting Question feature extraction and third sub-module for extracting External knowledge (fact) features.

Image feature extraction sub module:
Image features have been extracted using F-RCNN. We extract adaptive number of objects from each image maximum up to 100 objects. Given an image 'I', the extracted image features 'V' can be represented as,

$$Features(I) = V = [V_1, V_2, V_3, \ldots . V_K], V_i \in R^K \quad (1)$$

Where V_i represents the i^{th} object proposal feature in the image. 'K' is the number of objects regions (up to maximum K=100). And m = 2048 is the dimension of each object feature. Thus over all, image features are represented as, $V \in R^{Kxm}$, m =2048 and K is adaptive.

Question Feature extraction module:
Question features are extracted using BERT encoder. The BERT encoder contains multiple stacked attention modules that are based on the scaled-dot product attention [9]. Fig. 2 shows the self-attention mechanism used in BERT encoder.

The scaled dot product attention unit takes three vector inputs: queries, keys, and values, with dimensions dq, dk, and dv, respectively. For simplicity, we take $dq = dk = dv = d$. Three matrices are formed: Q, K and V containing all queries, keys and values together respectively. Then the attention weights are obtained as given in (1),

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

Here, Matrix Q and K are multiplied and this product is then divided by square root of d. The result is multiplied by matrix V followed by Softmax function applied to it. Given a question,

$$Q = [Q_1, Q_2, Q_3, \ldots . Q_L],\ Q_i \in R^L \quad (3)$$

where L is the number of words in a given sentence, we pre-process the questions as per the BERT requirements. Each question is added with [CLS] and [SEP] tokens at the beginning and at the end of the sentence respectively. We keep the question length fixed to 14 words. Smaller questions have been padded to make them length of 14 and larger questions are truncated to 14 without losing any important information. Attention masks have been generated for each question so that model can differentiate between actual tokens and padded tokens. We represent these pre-processed questions as $\hat{Q}$.

$$\hat{Q} = Pre - process\ (Q) = \left[\widehat{Q_1}, \widehat{Q_2}, \widehat{Q_3} \ldots, \widehat{Q_L}\right] \quad (4)$$

'L' is the number of words in a given sentence. These pre-processed question is then fed to the BERT_ENCODER where it is passed through the **G** number of scaled-dot – product attention modules stacked one after another. We used G=8.

$$Q_A = BERT\_ENCODER\ (\hat{Q}) = [S_1, S_2, \ldots . , S_N] \quad (5)$$

The output of the BERT encoder is self – attended question features, Q_A. These self-attended question features are later used to guide the image features and fact features.

Fact feature Extraction sub module:
In order to extract relevant facts, all the object labels from the image have been extracted. We use F-RCNN to extract

object labels. Out of these, top 5 objects have been used for extracting external knowledge (see Fig. 3).

Also, we extract important object related key words from the question. Let,

$$Objects\,(I) = O = [\,O_1, O_2\,, \ldots\,, O_K\,] \quad (6)$$

Where O_1, O_2, ... O_k are the object proposals and K is the number of object regions in a given image. We use ConceptNet as an external knowledge to extract the topmost

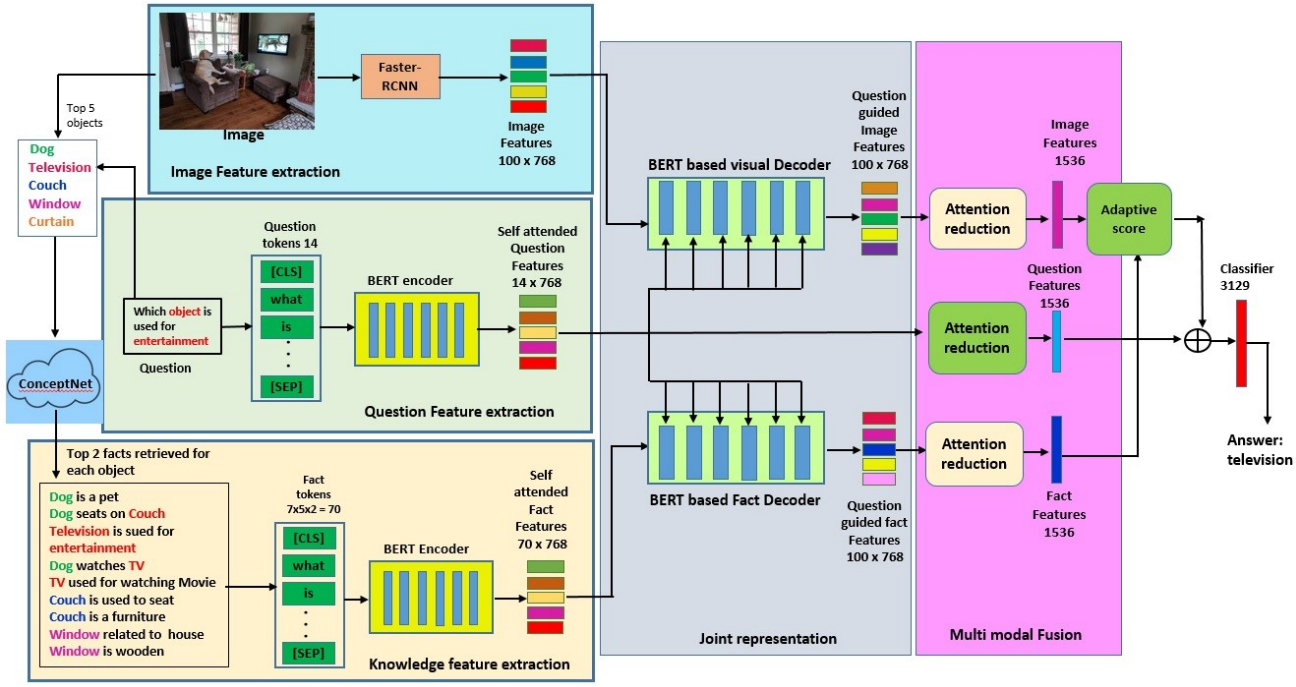

Fig. 1. Architecture of the Knowledge Blended Open-ended Visual Question Answering

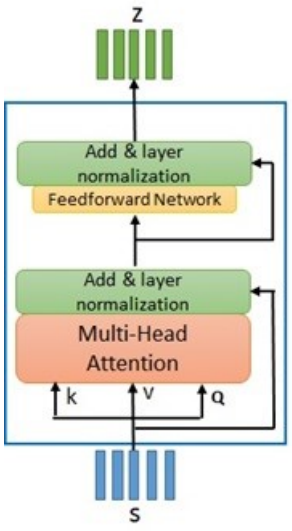

Fig.2. Scaled – dot product used in BERT encoder for self-attention (SA Unit)

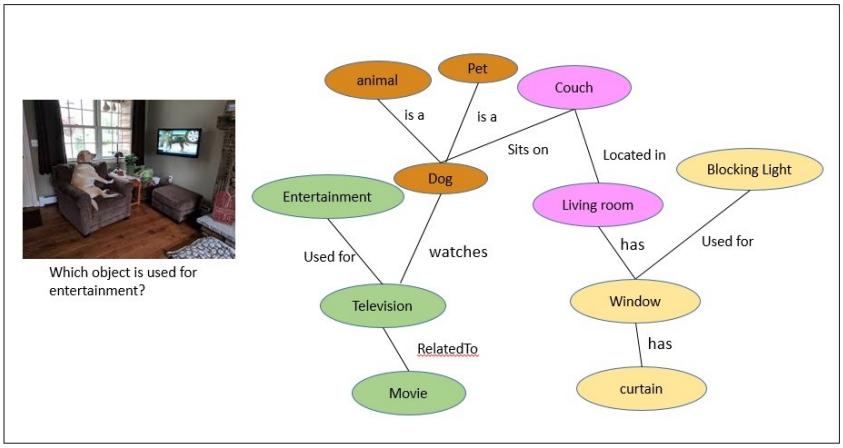

Fig. 3. Concepts extracted from the ConceptNet for the top 5 objects in an image and related keywords in a question.

two facts related to each of the top 5 objects. Each fact is of 7 tokens; we extract top 2 facts for 5 top objects. Thus the length of fact sequence for one image is 7x5x2 = 70 tokens. ConceptNet is publicly available structured knowledge base that contains the entities and the relationship between these entities. Fig. 3 shows an example of concepts extracted for a sample image.

The knowledge contains two components,

$$knowledge = \{fact, w\} \quad (7)$$

Thus, each fact is weighted by some weight w. More the weight, more relevant is the fact. And fact is given as,

$$fact = \{O,R\} \quad (8)$$

Where fact represents the knowledge and R represents the relationship against the object *O*. We weight each fact based on the similarity between the object and the relationship. Once we extract related facts, these facts cannot be directly used by the model. The concepts extracted from conceptNet are actually in the form of natural language sentences. We can represent each fact as,

$$f = [f_1, f_2, \dots, f_L] \quad (9)$$

where F_i is a single word in a given fact.

We need to convert these facts into lower dimension vectors. We pre-process these facts similar to questions and convert them into lower dimension vectors using BERT. Note that we consider only top 2 fact based on the weight w for further processing. Like questions, we use BERT encoder (BERT base with 12 layers) to extract fact features. The output of the BERT encoder is $\mathbf{F_A}$ i.e. self-attended fact features.

B. Joint Representation Module

Once the image features, self-attended question features (Q_A), and self-attended fact features (F_A) have been extracted, the important phase is to generate an effective attention mechanism to jointly comprehend these inputs: Image, question, and facts. The attention mechanism allows one modality to guide another modality. We use BERT-based decoder architecture for a multi-modal attention mechanism. Two decoders have been designed: Visual Decoder and Fact Decoder. Self-attended question features are used to guide both visual features and fact features.

1) Visual decoder: Visual decoder takes self-attended question features QA (generated from BERT encoder) and image features V (extracted using F-RCNN) and outputs question-guided image features VG.
2) Fact decoder: Fact decoder takes self-attended question features QA (generated from BERT encoder) and fact features FA (extracted using BERT Encoder) and outputs question-guided fact features FG.

The decoder contains **G** number of stacked Multi-modal Attention units (MMA) shown in Fig. 4. Each MMA unit takes two inputs: Input S and input (I OR F). For the visual decoder Image (I) is considered to be a query and a question (S) is considered to be key and value. The output of this decoder is the question attended visual features we call as V^G. Similarly, for Fact decoder, fact (F) is considered to be a query, and question (S) are keys and values. The output of this decoder is the question-guided facts, we call it F^G

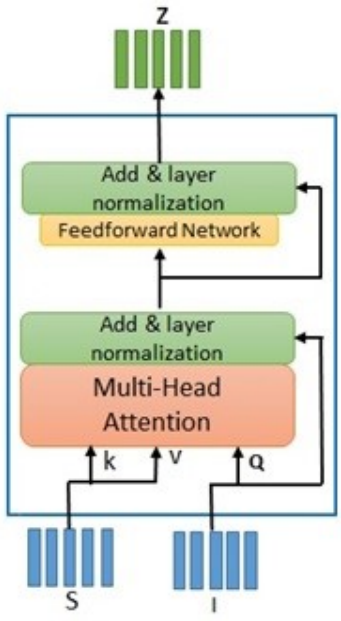

Fig 4: Multi-Modal attention unit (MMA)

C. Attention reduction module

After we obtain processed self-attended question features, decoded co-attended image features, and decoded co-attended fact features; these features are passed through the attention reduction module. This module is used to simply flatten the image, fact, and question features. The attention reduction module uses a simple two-layer MLP for its input.

D. Adaptive Score Generation Module

Not all questions require external knowledge. Some of the questions may be directly answered by visual clues and some questions may need external knowledge apart from visual clues. This completely depends on the question being asked. So, the job of the adaptive score generation module is to generate the score specific to an input question and automatically pick up one source of information that is more relevant to the question and treat another source as the sub-source. We compute the scores for the attended visual features V^G and the Attended Fact features F^G as follows.

$$Score(V) = W_2^V\ (\tanh(W_1^V\ V^G)) \quad (10)$$
$$Score(F) = W_2^F\ (\tanh(W_1^F\ F^G)) \quad (11)$$

Where $W_1^V \in R^{OXm}$, $W_2^V \in R^{OXO}$, $W_1^F \in R^{OXn}$, $W_2^F \in R^{OXO}$.

E. Answer generation Module

Adaptive score generation module help model identifying which features: visual or fact are more relevant to the given question. The output of adaptive score along with the self-attended question feature is given to the simple classifier (MLP) to generate a final answer. Due to the complexity involved in answer generation problem, we transform VQA problem into classification by considering top 3129 most frequently occurring answers as our final classes. The model

selects the best matches out of these classes. We also provide the soft score between [0,1] to each suggested answers.

IV. EXPERIMENTS

We trained our model on publicly available dataset VQA 2.0 [18] that contains approximately 2 lacks images taken from COCO image dataset, 1105904 questions and 11059040 ground truth answers. For each image there are three questions and for each question human generated 10 answers have been provided. The train, test and validation split is already given by the VQA authors [18]. We used kaggle colab for training the model. Model is trained for 13 epochs using adam optimizer, learning rate equal to 1e-4, $\beta1=0.9$ and $\beta2=0.98$ We used G=8 layers in BERT (BERT Medium) encoder and decoder with 512 hidden size, drop out rate 0.1 and batch size equal to 64. For evaluation we used the accuracy metric proposed in [1]. Since answers given by different humans may differ slightly, the following accuracy formula takes into account this human bias in answering a question.

Accuracy = min (Number of humans agreed with the model generated answer/3,1).

In other words, the answer is considered true if at least 3 humans said that answer.

V. RESULTS

We have evaluated our model on a validation set of the VQA 2.0 dataset. We compare two versions of our model. One without using an external knowledge base and one with an external knowledge base. We also vary, the number of layers in BERT encoder/decoder to observe its effect on model performance.

.

Table 1: Results of our model 'with the Knowledge base and without knowledge base on VQA 2.0 validation set.

Method	Number of Layers in BERT Encoder /Decoder	Open-Ended (validation accuracy)			
		Yes/No	Number	Other	All
Without KB	8	84.17	48.66	58.35	66.78
With KB	8	84.36	48.69	59.04	67.23
Without KB	12	84.87	48.31	58.66	67.15
With KB	12	84.90	48.72	59.15	67.97

Table 1 shows the accuracies of various proposed models on VQA 2.0 validation set. It is noteworthy that incorporating external knowledge increases the model accuracy as more complex open-domain answers can be correctly predicted. Note that the accuracy of the model in answering 'other' category questions has been increased. Also, changing the number of layers in BERT encoder/decoder affects model performance. As we increase the number of BERT layers in the encoder and decoder the model accuracy also increases.

Some of the results generated by our best model have been shown in Fig. 5

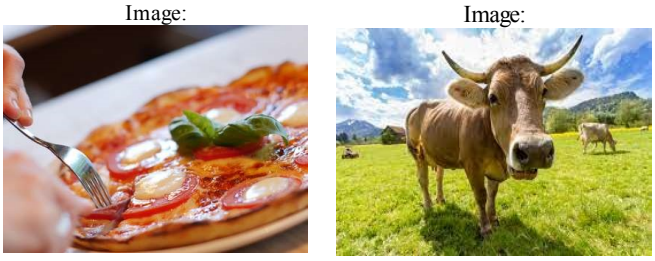

Fig. 5. Results of our model generated on random images

Table 2: Comparison of proposed method with other baseline models evaluated on VQA 2.0 validation set

Method	Open-Ended (validation accuracy)			
	Yes/No	Number	Other	All
Bottom-up attention [19]	80.3	42.87	55.81	63.2
N2NMN (2017) [20]	77.54	40.38	56.39	63.28
NSM (2019) [21]	79.77	41.75	59.40	65.77
XNM (2019) [22]	79.92	41.16	57.12	64.70
MRA-NET [23]	-	-	-	66.08
KI-NET [24]	81.92	43.16	60.47	67.32
OURS BEST MODEL WITH KB	84.90	48.72	59.15	67.97

Table 2 shows a comparison of our KB model with baseline models. Our experimental results show that the proposed model outperforms all the baseline models due to the incorporation of external knowledge from ConceptNet.

VI. CONCLUSION

We address the problem of answering open-domain questions that requires external knowledge in VQA model. The proposed model is a novel transformer based knowledge incorporated VQA model that is capable of answering open-domain questions. We successfully integrated the vast common sense knowledge available in

ConceptNet in the VQA frame work by utilizing the transformer as encoder to encode question and fact features. Also, power of transformer have been explored to perform joint comprehension between question and image features and between question and fact features. Out experimental results show that incorporating external knowledge improves model performance. The future work may include integrating multiple knowledge sources and take benefit of diversity of knowledge available in these Knowledge bases.

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] 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.

[3] K. Chen, J. Wang, L. Chen, H. Gao,W. Xu, and R. Nevatia,"ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering",. arXiv preprint arXiv:1511.05960, 2015

[4] J. Andreas, M. Rohrbach, T. Darrell and D. Klein, "Neural Module Networks," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 39-48, doi: 10.1109/CVPR.2016.12.

[5] C. Yang, M. Jiang, B. Jiang, W. Zhou and K. Li, "Co-Attention Network With Question Type for Visual Question Answering," in IEEE Access, vol. 7, pp. 40771-40781, 2019, doi: 10.1109/ACCESS.2019.2908035.

[6] Issey Masuda, Santiago de la Puente and Xavier Giro-i-Nieto. "Open-ended visual question answering." Bachelor's thesis, Universitat Politècnica de Catalunya, 2016.

[7] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra and D. Parikh, "Yin and Yang: Balancing and Answering Binary Visual Questions," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5014-5022, doi: 10.1109/CVPR.2016.542.

[8] K. J. Shih, S. Singh and D. Hoiem, "Where to Look: Focus Regions for Visual Question Answering," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4613-4621, doi: 10.1109/CVPR.2016.499.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin (2017). Attention Is All You Need. Advances in Neural Information Processing Systems. Curran Associates, Inc., 2017.

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

[11] N. Ruwa, Q. Mao, L. Wang and M. Dong, "Affective Visual Question Answering Network," 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2018, pp. 170-173, doi: 10.1109/MIPR.2018.00038.

[12] Guohao Li, Hang Su, & Wenwu Zhu,"Incorporating External Knowledge to Answer Open-Domain Visual Questions with Dynamic Memory Networks". CoRR abs/1712.00733.

[13] Q. Cao., B. Li,X. Liang & L Lin," Explainable High-order Visual Question Reasoning: A New Benchmark and Knowledge-routed Network", ArXiv, abs/1909.10128.

[14] Q. Wu, P. Wang, C. Shen, A. Dick and A. Hengel, "Ask Me Anything: Free-Form Visual Question Answering Based on Knowledge from External Sources," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4622-4630.

[15]] M. Ziaeefard and F. Lecue," Towards Knowledge-Augmented Visual Question Answering. In Proceedings of the 28th International Conference on Computational Linguistics", pages 1863–1873, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[16] Y. Zhang, M. Jiang and Q. Zhao, "Explicit Knowledge Incorporation for Visual Reasoning," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1356-1365, doi: 10.1109/CVPR46437.2021.00141.

[17] K. Marino, M. Rastegari, A. Farhadi and R. Mottaghi, "OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3190-3199, doi: 10.1109/CVPR.2019.00331.

[18] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra and D. Parikh, "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6325-6334, doi: 10.1109/CVPR.2017.670.

[19] P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077-6086, doi: 10.1109/CVPR.2018.00636.

[20] R. Hu, J. Andreas, M. Rohrbach, T. Darrell and K. Saenko, "Learning to Reason: End-to-End Module Networks for Visual Question Answering," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 804-813, doi: 10.1109/ICCV.2017.93.

[21] D. Hudson and C. Manning, "Learning by Abstraction: The Neural State Machine", 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, pages 5903–5916, 2019

[22] J. Shi, H. Zhang and J. Li, "Explainable and Explicit Visual Reasoning Over Scene Graphs," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8368-8376, doi: 10.1109/CVPR.2019.00857

[23] L. Peng, Y. Yang, Z. Wang, Z. Huang and H. T. Shen, "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, 1 Jan. 2022, doi: 10.1109/TPAMI.2020.3004830..

[24] P. Wang, Q. Wu, C. Shen, A. Dick, and A. Den Henge,"Explicit knowledge-based reasoning for visual question answering". Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17). AAAI Press, 1290–1296.