Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 31(4): 1948–1965.
DOI: 10.3934/era.2023100
Received: 08 December 2022
Revised: 11 January 2023
Accepted: 20 January 2023
Published: 13 February 2023

Research article

Visual Question Answering reasoning with external knowledge based on bimodal graph neural network

Zhenyu Yang[1,3], Lei Wu[2,*], Peian Wen[3] and Peng Chen[3,*]

[1] Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou 314099, China

[2] School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu 611731, China

[3] School of Computer and Software Engineering, Xihua University, Chengdu 610039, China

* **Correspondence:** Email: chenpeng@mail.xhu.edu.cn, wulei@uestc.edu.cn; Tel: +8613880092829, +8613084443881.

Abstract: Visual Question Answering (VQA) with external knowledge requires external knowledge and visual content to answer questions about images. The defect of existing VQA solutions is that they need to identify task-related information in the obtained pictures, questions, and knowledge graphs. It is necessary to properly fuse and embed the information between different modes identified, to reduce the noise and difficulty in cross-modality reasoning of VQA models. However, this process of rationally integrating information between different modes and joint reasoning to find relevant evidence to correctly predict the answer to the question still deserves further study. This paper proposes a bimodal Graph Neural Network model combining pre-trained Language Models and Knowledge Graphs (BIGNN-LM-KG). Researchers built the concepts graph by the images and questions concepts separately. In constructing the concept graph, we used the combined reasoning advantages of LM+KG. Specifically, use KG to jointly infer the images and question entity concepts to build a concept graph. Use LM to calculate the correlation score to screen the nodes and paths of the concept graph. Then, we form a visual graph from the visual and spatial features of the filtered image entities. We use the improved GNN to learn the representation of the two graphs and to predict the most likely answer by fusing the information of two different modality graphs using a modality fusion GNN. On the common dataset of VQA, the model we proposed obtains good experiment results. It also verifies the validity of each component in the model and the interpretability of the model.

Keywords: visual question answering; external knowledge; bimodal fusion; pre-trained language models; knowledge graphs

1. Introduction

VQA is an attractive research direction [1], which aims to analyze multimodal content from images and questions. VQA model has grounded, reasoning, and translation capabilities and can answer questions in natural language based on images. In recent works [2–8], VQA problems, which refer only to visible image content and only use local computational resources to solve VQA tasks, have been very successful. As shown in Figure 1, considering the problem, the model needs to visually locate the "fruit" and relate the knowledge that "banana is sweet and healthy". Thus, collecting evidence and additional information related to the problem from different models is critical to achieving broad VQA; on the other hand, after information gathering is complete, the VQA system must combine the information obtained to infer and obtain the answer to the question.

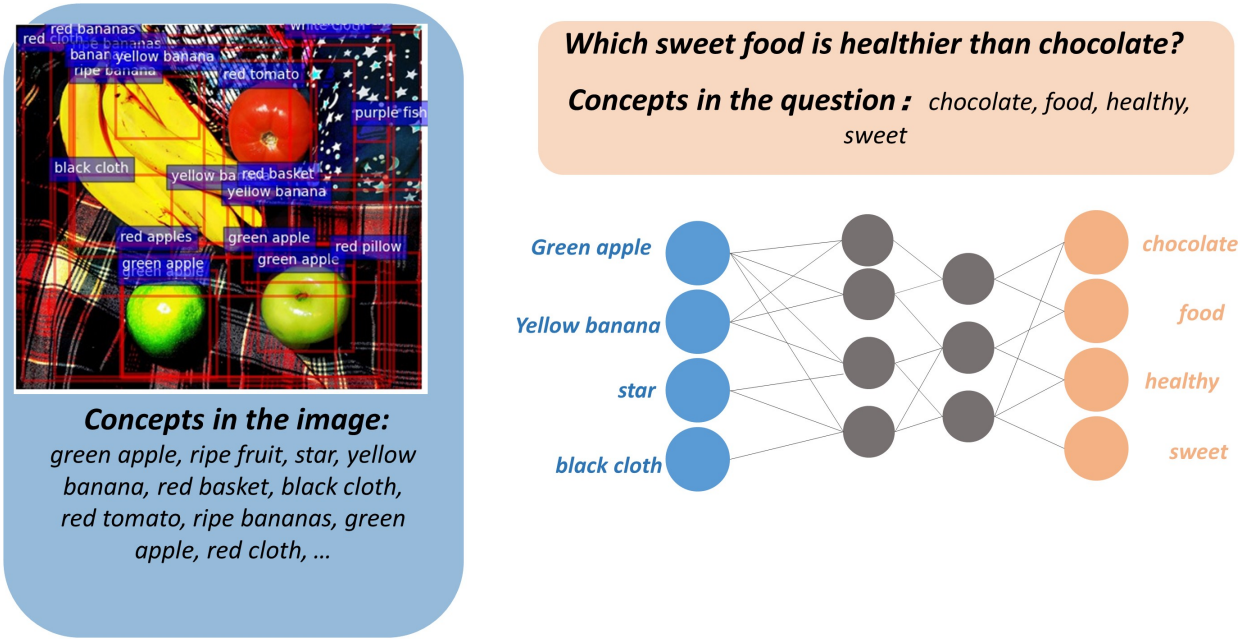

Figure 1. The example of VQA task.

Existing work [9, 10] solves the problem by trans the questions to keywords and retrieving supporting entities only through keyword matching to get the answers. However, the proposed method is vulnerable to attack when the problem does not accurately mention visual concepts (such as synonyms and homographs). The information referred to is not captured in the graph of the facts (for example, the visual concept "yellow" in Figure 1 may omission will cause errors). This method can efficiently reduce the amount of computation compared to the traditional semantic similarity-based reasoning method. In order to solve these problems, [11] introduced visual information to the actual graph and used the implicit knowledge graph to infer the answer based on problem-based reasoning.

On the other hand, although implicit knowledge LMs have a broad range of knowledge coverage, they could improve in terms of experience with structured reasoning [12]. In contrast, KGs are more suited to structured reasoning [13, 14]. They can obtain interpretable predictions by providing paths for reasoning [15] but can lack coverage and be noisy [16, 17]. Thus, the LM+KG module, combining the

complementary benefits of LMs and KGs, the LM+KG modal may improve the reasoning performance of the VQA algorithm and the prediction accuracy of the answers to the data set.

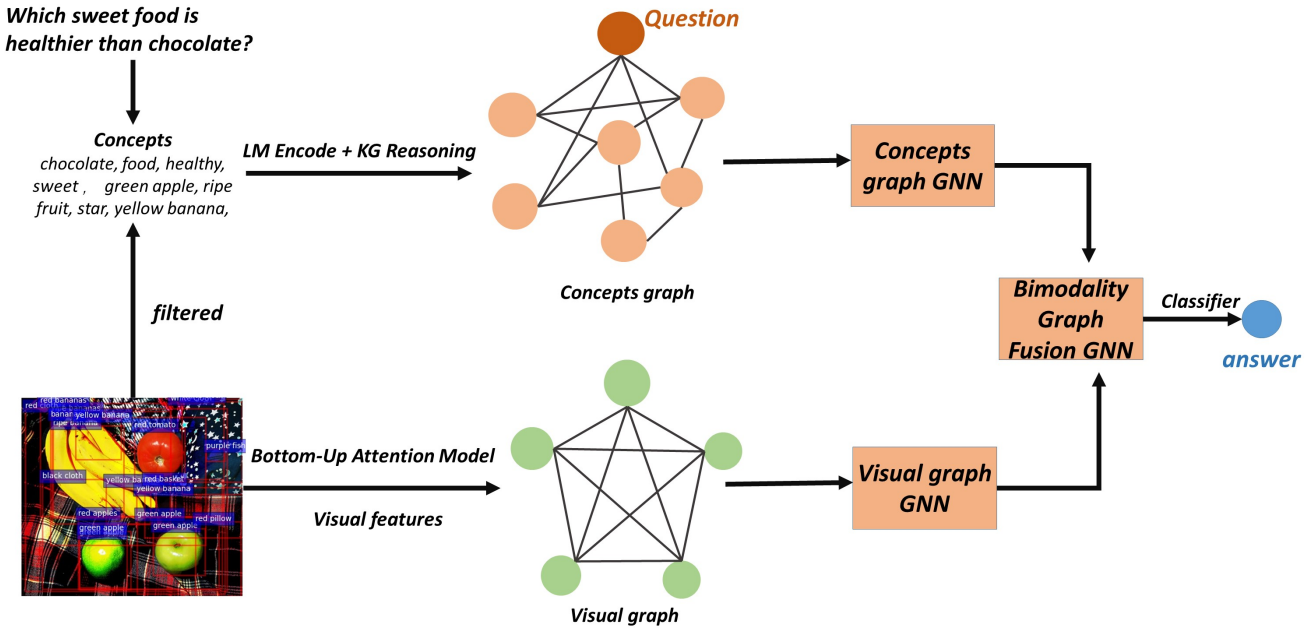

Figure 2. The framework of model BIGNN-LM-KG.

This article introduces a model that co-learns from vision, language, and KG embedding and captures specific interactions in images and problem concepts. Our method BIGNN-LM-KG flow in Figure 2. First, we will extract the conceptual information in the pictures and questions and use pre-trained LM to encode the extracted in the problems and image entity concepts. We use two graph types to construct the problems and images given by VQA tasks. We use the Kagnet [18] to filter nodes and edges effectively in the constructed graph. We specially constructed the concept graph of the existing concepts using KG reasoning. At the same time, we use entity features and location information to construct a visual graph for the information contained in the images. Then, the concepts and visual graphs are sent into the improved GNN to learn the representation. The representation learned by these two kinds of graphs uses a modal fuse GNN for knowledge integration between modalities. Finally, the integration results are fed to the classifier to predicate the answer. In this article, our main contributions include the following aspects:

- For integrating the two information modalities, the difference between conceptual and visual information makes integrating the two information modalities. Here we propose a GNN neural network based on modal perception to extract evidence related to question answers from two different modalities. Notably, this paper uses the modality fusion GNN to filter out the concept information with the highest degree of relevance based on the modal information gained from the fusion and then perform the reasoning of the answer based on this.
- For it is difficult to deduce valid information if the VQA system constructs a conceptual diagram directly from conceptual information extracted from pictures and problem text to predict the answer. Therefore, we use LM + KG's joint inference advantage to construct the graph and add

node correlation scores to the conceptual graph to generalize the weight of KG information. So that makes the constructed conceptual diagram contains information that is more relevant to the answer to the question.

The organizational structure of this paper is as follows: The first section mainly introduces the research background and purpose of VQA and briefly analyzes the main problems faced by VQA at present. The second section reviews the classical models in the VQA field and introduces the development of Graph Neural Networks (GNN), LM, and KG commonly used in VQA. The third section briefly introduces the VQA model BIGNN-LM-KG proposed in this paper and introduces the construction of different modality graphs and their fusion methods. The fourth section verifies the effect of the model in the VQA dataset, conducts an ablation study on various parts of the model to verify the help of different parts of the model for the whole, and demonstrates the interpretability of the model. The fifth section discusses the advantages and disadvantages of the proposed model. The last section summarizes the article and prospects the further research in the VQA field.

2. Related work

2.1. Knowledge-based VQA

The CNN-RNN architecture using global visual features to represent images is a typical solution of VQA [18]. In the following years, researchers have introduced attention mechanisms [19, 20] and relational reasoning mechanisms [21–23] to improve VQA accuracy continuously. However, this process has ignored one of the most critical points. When answering VQA questions, people will involuntarily combine external knowledge and visual information, but it is not easy for the VQA algorithm. Therefore, [24] proposed a new VQA dataset named FVQA. The questions in this dataset introduce supporting fact associations; the VQA model must reason about various facts when answering questions.

The work based on FVQA usually requires selecting an entity from the facts as the answer. The method in [10] can infer the predicted answer according to the information extracted from the knowledge base and pictures. However, its approach relies predominantly on predefined templates. So it has stringent requirements on the format and type of questions. At the same time, this method is only used to extract entities from image information without contacting the content entities of the questions when making inferences in the prediction of answers. In [25], proposed adding visual relationship judgment between different objects and a problem-oriented attention mechanism, but there were great difficulties in relationship judgment during the experiment. In [26], proposed multi-scale relational reasoning to conduct multimodal VQA and designed a regional attention scheme to help extract information and regions of interest related to the problem, which achieved good results in VQA data sets without external knowledge. This paper extracts conceptual information from the picture and the question text and combines external knowledge. Moreover, use LM+KG joint reasoning to construct the concept graph, obtain evidence related to the prediction problem, and predict the problem's solution by studying the multi-layer GNN network structure.

2.2. Graph neural networks

In recent years, GNN has been developing rapidly [27]. Although isomorphic graphs have many applications, in reality, most of the graphs are heterogeneous. The primary graph convolution network (GCN) [28] only uses the neighborhood information of the graph for message transmission and cannot distinguish all relationship types. Relation-aware Graph Convolution Network (RGCN) [29] popularizes GNN by coding each side with different relations separately to fit the multi-relational graph, then dealing with the knowledge base is different relations between entities. Heterogeneous Graph Attention Networks (HAN) [30] and Heterogeneous Graph Attention Networks for Semi-supervised Short Text Classification (HGAT) [31] developed a heterogeneous graph focus network with a two-layer focus mechanism. These methods are all used to model different nodes and edges in the unified graph. In the research work of this paper, we build a heterogeneous graph containing different modal information to realize the fusion and reasoning of the information contained in VQA. After the representation learning through the intra-modal convolution of the GNN network, we also use the modality fusion GNN to conduct cross-modal convolution reasoning of the learned representation to obtain the final answer prediction.

2.3. LM+KG

The large-scale knowledge base has become an essential external knowledge representation resource by organizing the world's facts into structured databases. A set of object triples (also called facts) of the subject-predicate is formed in a typical knowledge base. This type of knowledge base is often called a Knowledge Graph (KG) [32] due to its graphical representation. The entity is the node, and the relationship is the edge of the linked node. In a triple, two entities are specified as being related by a specific relation, such as $< Biden, America, president >$.

Despite the remarkable success of pre-trained large language models LMs in many QA tasks [33, 34], there is a paucity of work on their integration with KG and image representation in VQA tasks. A new language representation method, the BERT-based token embedding method, is proposed by [35]; however, this model is also a query-based approach. The algorithm first takes an entity name as a node and inserts a corresponding KG triplet on it to obtain a new node; The node then injects entities into the query to solve the problem.

Generally, knowledge can be implicitly encoded in pre-trained LM on unstructured texts or explicitly expressed in structured KG. All these two methods are widely used in the field of natural language processing. Moreover, recent research has also been devoted to the combination of LMs and KG [36, 37]; the methods of these articles retrieve subgraphs on KG by obtaining subject entities, that is, KG entities and their multi-hop neighbors mentioned in the given information. However, this introduces many entity nodes unrelated to the given information semantics. These nodes introduce not only different subject entities but also multiple hops. In this paper, we use LM+KG to jointly infer the concept information extracted from the VQA task to build the concept graph. Specifically, we use KG to match the relationship subgraph between the entity concepts extracted. LM will encode the nodes and edges in the subgraph to form a concept graph.

3. Methods

For the prediction of VQA task answer A with external knowledge, in addition to the question Q and image I provided by itself, we also need a knowledge base containing facts in the form of triples, namely $< c1, r, c2 >$, where $c1$ and $c2$ are entity concepts extracted from images and questions. R represents the relationship between $c1$ and $c2$. On this basis, this paper proposes a VQA problem-solving framework based on external knowledge. This framework can select the objects that best meet the user's needs from multiple supporting entities, classify these objects, and then give corresponding $c1$ or $c2$ prediction answers according to the characteristics of different categories.

First, we need to extract the conceptual information from the picture and question text and then use KG to match the relationship subgraph between the entity concepts extracted. LM will encode the nodes and edges in the subgraph to form a concept graph. At the same time, the spatial and visual information in the image will be composed to form a visual graph. On this basis, we propose a reasoning method based on a bimodal fusion GNN. It selects the knowledge associated with the question from each layer of the graph through the convolution of the inner modality graph from a single mode. Performs iterative reasoning from two modes adaptively through the convolution of modal fusion graph and obtains the answer with the highest probability through joint analysis of entities. The picture shows the detailed structure of our model.

3.1. Identify the concept of task inclusion

A simple method of concept recognition is to precisely match the entity concept in the sentence with the existing nodes in the KG. For example, in the question "Which sweet food is healthier than chocolate?" the exact matching results QC will be fitting, *chocolate*, $food$, *healthier*, *sweet*. We will use some rules in Kagnet [10] to improve this direct method, such as soft matching via lemmatization and stop word filtering. According to the concept identification results of the question text, we will name the remaining K concepts QC_i.

Given the image *Img*, we use the Bottom-Up Attention Model [38] to identify the entity conceptual objects in a group of images. $FC = [FC_1, FC_2, ..., FC_{36}]$, where each object FC_i includes the visual feature vector CV_i ($d_v = 2048$), spatial eigenvector CB_i ($d_b = 4$) and associated entity label IC_i. Specifically, $CB_I = [x_i; y_i; d_i; h_i]$, where (x_i, y_i) represents the coordinates of the upper left corner, and h_i and d_i represent the height and width of the bounding box, respectively. For each entity concept IC_i in the picture. The concept group IC comprises Ic_i, and the concept group result QC is extracted from the question text through the Kagenet method. Use the pre-trained LM text vector model to vectorize the IC and QC and calculate the Word Mover Distance for the vectorized two vector groups. The calculation process is as follows:

$$\min_{T \geqslant 0} \sum_{i,j=1}^{n} T_{i,j} \cdot d(i, j) \tag{3.1}$$

$$\sum_{j=1}^{n} T_{i,j} = QC_i \qquad i \in (1, ..., k) \tag{3.2}$$

$$\sum_{j=1}^{n} T_{i,j} = IC_j \qquad j \in (1, ..., k) \tag{3.3}$$

where $T \in \mathbb{R}^{n+n}, T_{i,j} \geqslant 0$ indicates the word QC_i in vector group QC transfer to another vector group IC Word IC_i distance weight. The $d\,(i, j)$ is the distance of the word between QC_i and IC_i.

After the image concept group is filtered, only the first k distance image concepts IC_i between the text vector QC have been retained. Each object's visual and spatial characteristics correspond to each concept.

3.2. Graph construction

3.2.1. Entity concept graph

A practical method to effectively match different concepts of question-answer pairs in QA questions is proposed in Kagnet [18]. This paper also adopts this method for each question concept $QC_i \in QC$ and image entity concept $IC_i \in IC$. We can effectively find the path between concepts shorter than k-hop [15]. In order to use the knowledge obtained from LM and KG to infer the given entity concept, this paper uses the pre-trained LM to obtain the entity concept expression in the question and image. And then adds an edge between the concept pairs in QC or IC to retrieve that there is a joint concept subgraph $G_{Concept}$ in KG. This subgraph contains two knowledge sources: questions and pictures.

Further, to preserve the connection between the question text and the nodes of the joint concept subgraph, this paper uses the research method of QAGNN [39] to introduce a new question text node Qz on the joint concept subgraph. The node Qz is obtained by LM reasoning the question text. Qz and the subject entities on $G_{Concept}$ are connected.

Each node in $G_{Concept}$ belongs to the following four types: the context node QT, the node in QC, the node in IC, and other nodes introduced from KG. Many nodes on the KG subgraph $G_{Concept}$, that is, nodes retrieved from the KG heuristic, can interfere with the prediction of the current question answer. As can see from the retrieved KG subgraph $G_{Concept}$ in Figure 3, the k-hop neighbor in QC may contain some nodes that are not beneficial to the reasoning process, such as the node "holiday" and "riverbank" deviate from the theme; "People" and "local" are ordinary. These irrelevant nodes will cause overfitting or unnecessary noise to reasoning, especially when there are many QC and IC nodes. Taking ConceptNet [40] as an example, even if only 3-hop neighbors are considered, $G_{Concept}$ will become a KG subgraph with an average of more than 400 nodes.

To reduce the $G_{Concept}$ number of nodes and paths. This paper uses a node correlation score. In this method, the correlation score of each node Cn on the KG subgraph $G_{Concept}$ is calculated through the pre-trained LM, and the basis for scoring is the question text Q provided by the task. For $G_{Concept}$, each node Cn connects the entity concept text $text(Cn)$ with the question text $text(Q)$ to calculate the correlation score Cs:

$$Cs = LM_{head}(LM_{enc}([text(C_n)]; text(Q))) \tag{3.4}$$

where $LM_{head} \circ LM_{enc}$ represents the possibility of the $text(Cn)$ of the entity concept calculated by LM, the correlation score C_s is the importance of each node Cn in $G_{Concept}$. It is relative to the given question Q helps $G_{Concept}$ prunes the path and delete the node.

Specifically, for subgraph $G_{Concept}$ after path pruning and node deletion, the embedding t_u of node u and the embedding r_uv of the relationship between node u and node v are defined as follows:

$$t_u = W_1 t_u + W_2 \tag{3.5}$$

$$r_{u,v} = MLP(W_3(e_{u,v}, t_u, t_v)) \tag{3.6}$$

where $W_1, W_2, ..., W_{23}$ are learning parameters, t_u,t_v is the representation vector of the text of node u, v, e_{uv} is the representation vector of the relationship type between u and v nodes.

3.2.2. Image visual graph

After screening the image concept group, we retain the first k image concept labels IC related to question Q and the visual and spatial characteristics of each object entity corresponding to each concept label. Based on the existing visual and spatial features, we use the concept label IC_i corresponding entity object feature IC_F features as the graph node. While the spatial feature vector $Fb_i = [x_i; y_i; d_i; h_i]$. After the following encoding, it is used as the edge $s_{i,j}$ of the graph:

$$s_{i,j} = \left[\frac{x_j - x_i}{d_i}, \frac{y_j - y_i}{h_i}, \frac{d_j}{d_i}, \frac{h_j}{h_i}, \frac{d_j \cdot h_i}{d_i \cdot h_i}\right] \tag{3.7}$$

The visual graph and concepts graph construction overview will show in Figure 3.

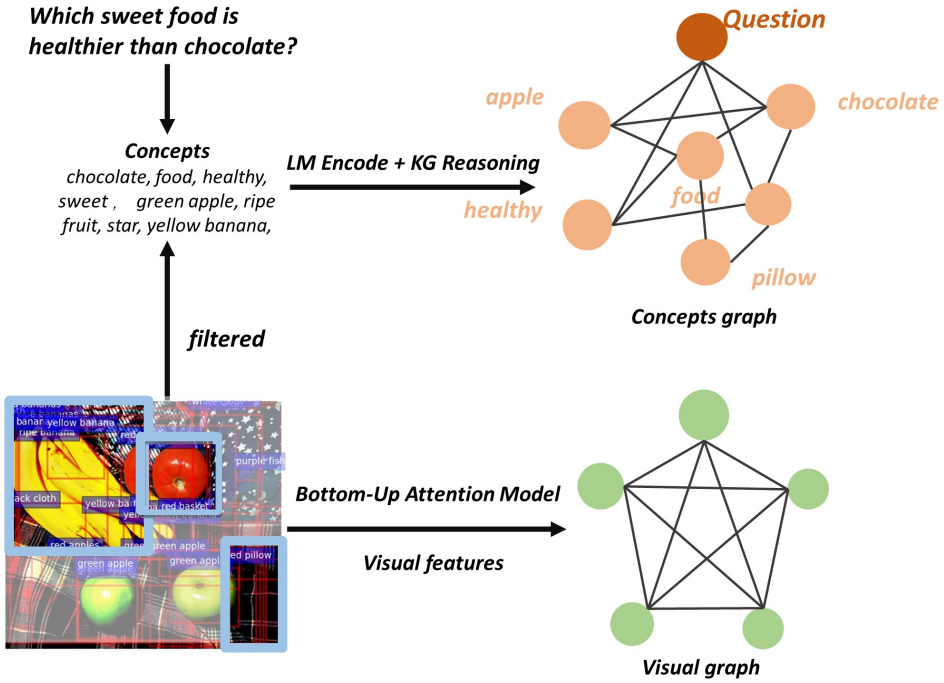

Figure 3. The visual graph and concepts graph construction overview.

3.3. GNN construction

Because the concept and visual graphs contain specific knowledge of various modalities related to the question, this paper makes some improvements. It is based on GAN [41] according to the characteristics of the visual graph and the concept graph so that it can independently filter out the evidence of significance from these two graphs to predict the answer to the question.

3.3.1. Concept graph GNN

This paper establishes a GNN architecture implementing the graph $G_{Concept}$ of reasoning based on the graph attention framework GAT [30] for the concept graph module. The update of node representation in this model comes from the message passing between adjacent nodes on the graph. Specifically, for an n-layer GNN, in each layer, GNN will update the representation $U_p^{(n)} \in \mathbb{R}^D$ of each node p in the following way:

$$U_p^{(n+1)} = RELU\left(W_4\left(\sum_{u\in\mathcal{N}_p\cup\{v\}}^{n} \alpha_{u,v} + m_{u,v}\right)\right) + W_5 U_p^{(n)} \tag{3.8}$$

where $\mathcal{N}_p$ represents the neighborhood of node p, $m_{u,v}$ represents the message in the layer passed from node u to neighbor node v, $\alpha_{u,v}$ indicates the attention weight when the message $m_{u,v}$ is passed from u to v. Moreover, $m_{u,v}$ will be calculated according to the following calculation process.

Relationship-aware messages. The message passing from two nodes of the multi-graph $G_{Concept}$ needs to include the relationship between the two nodes. Therefore, in order to facilitate the calculation of message passing, the message from node u to node v is:

$$m_{u,v} = W_6 U_p^{(n)} + W_7\left[t_u, r_{u,v}\right] \tag{3.9}$$

For the vectorization of the correlation score of node p, we will use the following formula to determine:

$$Cs_p = MLP(W_8 Cs_p) \tag{3.10}$$

For the D dimension attention weight between node u and node v:

$$q_u = W_9[U_u^{(n)}, [t_u, Cs_u]] \tag{3.11}$$

$$k_{u,v} = W_{10} U_v^{(n)} + W_{11}[t_v, Cs_v, r_{u,v}] \tag{3.12}$$

$$\alpha_{u,v} = \frac{\exp(\frac{{q_u}^{\mathrm{T}} + k_{u,v}}{\sqrt{D}})}{\sum_{v'\in\mathcal{N}_p\cup\{u\}} \exp(\frac{{q_u}^{\mathrm{T}} + k_{u,v}}{\sqrt{D}})} \tag{3.13}$$

3.3.2. Visual graph GNN

For a GNN of n layers in the visual graph, the GNN will update the representation of each node p of the visual graph in the following way for each layer:

$$\widehat{U}_p^{(n)} = RELU\left(W_{12}\left(m_p, \alpha_p, \widehat{U}_p^{(n+1)}\right)\right) \tag{3.14}$$

For each node in the visual graph and each edge between two nodes, we calculate the attention weight of node u and the attention weight of the edge $s_{u,v}$ between node u and node v through its associated question text Q:

$$\alpha_u = softmax\left({W_{13}}^{\mathrm{T}} \tanh\left(W_{14}F_u + W_{15}q\right)\right) \tag{3.15}$$

$$\beta_{u,v} = softmax\left({W_{16}}^{\mathrm{T}} \tanh\left(W_{17}F_v{}' + W_{18}q'\right)\right) \tag{3.16}$$

where q represents the quetion Q embedding vector obtained from LM,F_u represents the entity features of node u , $F_v{}' = W_{19}[F_v, s_{u,v}]$, $q' = W_{20}[F_v, q]$, $and[\cdot,\cdot]$ represent concatenation operation.

Based on the above Eqs (14)–(16), the message passing from node u to node v is:

$$m_{u,v} = \sum_{u \in N_p \cup \{v\}} \beta_{u,v} F_v{}' \tag{3.17}$$

3.3.3. Bimodal graph fusion GNN

The answer to the VQA model usually comes from the entities on two modality graphs. This paper uses the bimodal fusion GNN to collect the complementary information on the visual graph and fuse it with the concept graph. Finally, after the fusion, the judgment is made to infer all entities to form a global decision on the answers.

$$\gamma_{u,v} = softmax\left({W_{21}}^{\mathrm{T}} \tanh\left(W_{22}\widehat{U}_u^{(n)} + W_{23}\left[U_v^{(n)}, q\right]\right)\right) \tag{3.18}$$

where $\gamma_{u,v}$ represents the attention weight between the concept and visual graphs. So, the message passing of Bimodal graph fusion GNN is:

$${m_{u,v}}^{V-C} = \sum_{u \in N_p \cup \{v\}} \gamma_{u,v} \widehat{U}_v^{(n)} \tag{3.19}$$

Through the bimodal fusion GNN, we try to integrate valuable information from concepts and visual graphs. The fusion information will be iterated continuously to obtain the final entity representation of the VQA task answer. Then we input these entity representation and question representation vectors into the binary classifier and select the entity with the highest probability of answering from the entities for output.

4. Experiments and results

This section will describe the experimental environment, the datasets used, and the evaluation metrics used to evaluate the models we present.

4.1. Implementation details

We set the size of the GNN module (D = 300) and the number of layers (L = 3), and the dropout rate of each layer is 0.3. When we train the model, we use the Adam optimizer, which has 30 epochs, the batch size is 32, and the learning rate for the module is 3e-4.

4.2. Public datasets

We used two VQA datasets, FVQA and Visual7W+KB, to verify the effect of the model. We use the more classic CP knowledge graph as a supplement for external knowledge.

ConceptNet. Knowledge graph ConceptNet, a general domain knowledge graph, serves as our structured knowledge source graph [40]. There are 799,273 nodes and 2,487,810 edges in this knowledge graph. The node embedding is initialized by the entity embedding written in Kagnet. The entity is embedded on the ConceptNet to apply the pre-trained LMs to all triples. Then the set representation of each entity is obtained.

FVQA. FVQA dataset [10] includes a knowledge base of 2190 pictures, 5286 questions, and 193449 facts. The images in the dataset are randomly collected from the MSCOCO [42], and the original 80k-40k training and verification segmentation is used as the training and test segmentation. The fact is constructed by extracting top-level visual concepts from image data sets and querying them on WebChild, ConceptNet, DBPedia, and other KGs.

Visual7W+KB. Visual7W dataset [43] is based on a subset of images selected from the visual genome [44], which includes multiple questions and answers in the form of multiple choices. In addition, Visual7W [45] created a set of knowledge-based test questions by filling in a question-and-answer mode (KB dataset) composed of visual content and external knowledge. Name Visual7W+KB uses ConceptNet to manage issues but does not provide a knowledge base for specific tasks.

4.3. Experimental results

The model proposed in this paper is compared with the existing model On the FVQA dataset and the Visual7W+KB dataset. Furthermore, the results are shown in Tables 1 and 2.

Table 1. Experimental result on FVQA dataset.

Method	Overall Accuracy	
	Top-1	Top-3
LSTM-Question+Image+Pre-VQA	24.98	40.4
Hie-Question+Image+Pre-VQA	43.14	59.44
FVQA (top-3-QQgraphing)	56.91	64.65
FVQA (Ensemble)	58.76	-
Straight to the Facts (STTF)	62.2	75.6
Out of the Box (OB)	69.35	80.25
BIGNN-LM-KG(ours)	71.27	83.59

Table 2. Experimental result on Visual7W+KB dataset.

Method	Overall Accuracy	
	Top-1	Top-3
KDMN-NoKnowledge	45.1	-
KDMN-NoMemory	51.9	-
KDMN	57.9	-
KDMN-Ensemble	60.9	-
Out of the Box (OB)	57.32	71.61
BIGNN-LM-KG(ours)	67.59	79.84

It can be seen that compared with the best model OB in the FVQA dataset, the model in this paper improves the Top-1 accuracy by 1.92% and improves the Top-3 accuracy by 3.34%. The model in this paper, in the Visual7W+KB dataset, has improved 10.27% in Top-1 accuracy and improved 8.23% in Top-3 accuracy compared with the best model OB. Model OB is similar to the research idea of our model, and it also uses a GNN to evaluate all entities. However, it introduces entities equally without selection or screening. The model in this paper has been significantly improved by using LM+KG to construct the concept graph and introducing the GNN of the modality fusion.

4.4. Ablation study

In this section, we present the results of ablation experiments to verify the contributions of various components in the model. As for the LM+KG modal, it can be seen from Table 3 that for a complete model, if the model deletes the LM+KG part, it only uses the question concept and image entity concept to construct graphs without any joint reasoning among knowledge. It is easy to find that after comparing with the complete model, the TOP-1 accuracy without the LM+KG part model structure will decrease by 3.1%. It shows that the problem concept, image entity concept extraction, and related part processing can provide some basis for the model inference for VQA tasks.

Table 3. Ablation study result on Visual7W+KB dataset.

Method	Overall Accuracy	
	Top-1	Top-3
w/o LM+KG modal	64.49	74.62
w/o Modality fusion GNN	67.42	78.46
BIGNN-LM-KG((full model))	67.59	79.84

For the part of modality fusion GNN, the role of the visual graph in the VQA task has been verified many times. Therefore, compared with the complete model, we only delete the modality fusion GNN used in the model for the fusion and selection of the concept graph and visual graph. As seen from the table, the answer accuracy of the model has decreased by 1.9% in TOP-1. It shows that the knowledge representation for different modalities must be fused to use cross modalities knowledge effectively. However, concept graphs have a more significant impact on the accuracy of answers. The lack of

reasoning knowledge models can not better predict answers.

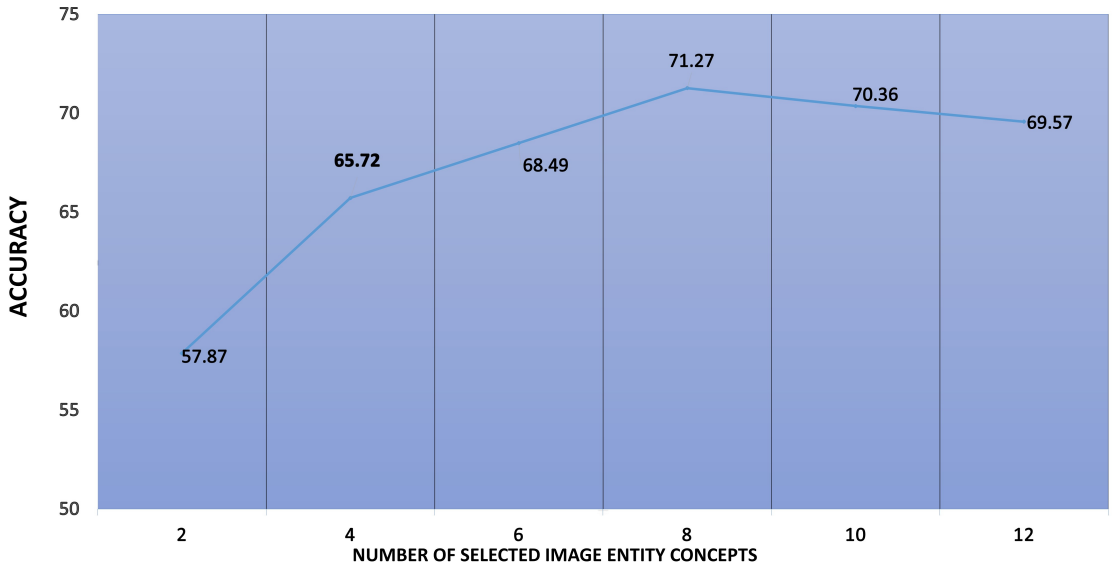

Figure 4. This figure shows the influence of the number of selected image entity concepts on the accuracy of the model.

For selecting visual concepts in images, we studied the influence of the number of graphic entity concepts K on visual question answering. In Figure 4, the increase of K in the early stage improved the accuracy of the answer until $K = 8$. However, after $K = 12$, the processing performance of the model began to decline, and the accuracy of the model's prediction response began to decline. The reason may be that the more entity concepts selected, the more paths and nodes in the concept graph, and the more nodes and longer relationship paths, the greater the noise.

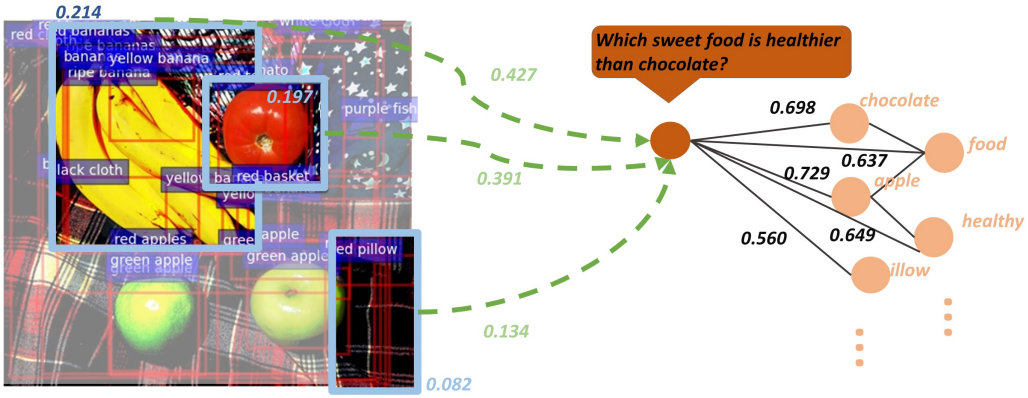

Figure 5. Visualization results of partial weights in the BIGNN-LM-KG model.

5. Discussion

Attention weight and correlation score can better explain our model in the visual reasoning process. Through the example in Figure 5, we can have the following insights:

1) BIGNN-LM-KG can reveal the importance of different modalities of information to the problem. It can be seen from Figure 5 that BIGNN-LM-KG can give rational correlation scores and attention weights for all the information given by VQA tasks. From the threshold values given by these models, it can be found that concept graphs can provide more essential hints than visual graphs. It is not only that the concept graph contains entity concepts such as questions, images, and KG. However, external knowledge is more attention to when building FVQA datasets.
2) Using the model for bimodality fusion, GNN can better promote the information interaction between modalities. The model not only gives the attention weight value or correlation score related to the problem for the nodes between single graphs. However, it also provides attention to the weight value between the nodes with a high correlation between the two charts, which can directly explain the mutual influence of information between different modes, indicating that it is necessary to fuse modalities.
3) Despite the provision of good correlation scores and attention weights for each node in the concept graph, Figure 5 clearly shows that for similar conceptual entities, the correlation scores and attention weights between the nodes are relatively similar. The VQA may get a relatively close but still incorrect answer when it answers questions.

6. Conclusions

This paper proposes BIGNN-LM-KG, a VQA task model for introducing external knowledge. It focuses on the joint reasoning of the question text concept and the selective picture entity concept.

We propose the KG joint reasoning based on the question text concept and the selective picture entity concept. We linked the picture, question text, and KG information source through the concept graph, embedded the concept, and generated correlation scores through LM. On this basis, the visual image is constructed from the filtered entity space and visual features in the image. The two graphs' representation is updated through GNN. The different modalities in the concept graph and visual graph are combined by the modal fuse GNN network so that the research model can more efficiently use the information between different modes to solve the same task jointly.

The model BIGNN-LM-KG we have built has achieved ideal results on various data sets, which is obviously superior to the frontier method. At the same time, we have also conducted interpretable experiments on the benchmark data set to verify the model's interpretability.

For future research, we will further improve the use and reasoning of visual information in images to ensure that various information provided by VQA tasks can be fully used, thus helping to improve the accuracy of answers. At the same time, the current research work is relatively simple for information fusion between modalities. In the next step, we will propose more practical and convenient methods for modality fusion.

Acknowledgments

This research was funded by the Municipal Government of Quzhou under Grant No. 2021D007, No. 2021D008, No. 2021D015, No. 2021D018, and No. 2022D029; Science and Technology Program of Sichuan Province under Grant No. 2021JDR0222 and No. 2020YFG0326 and Talent Program of Xihua University under Grant No. Z202047 and No. Z222001.

Conflict of interest

The authors declare there is no conflicts of interest.

References

1. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, et al., Vqa: visual question answering, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2015), 2425–2433. https://doi.org/10.1109/ICCV.2015.279
2. R. Cadene, H. Ben-Younes, M. Cord, N. Thome, Murel: multimodal relational reasoning for visual question answering, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 1989–1998. https://doi.org/10.1109/CVPR.2019.00209
3. L. Li, Z. Gan, Y. Cheng, J. Liu, Relation-aware graph attention network for visual question answering, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 10313–10322. https://doi.org/10.1109/ICCV.2019.01041
4. H. Ben-Younes, R. Cadene, N. Thome, M. Cord, Block: bilinear superdiagonal fusion for visual question answering and visual relationship detection, in *Proceedings of the AAAI Conference on Artificial Intelligence (AI)*, **33** (2019), 8102–8109. https://doi.org/10.1609/aaai.v33i01.33018102
5. C. Song, M. Liu, J. Cao, Y. Zheng, H. Gong, G. Chen, Maximizing network lifetime based on transmission range adjustment in wireless sensor networks, *Comput. Commun.*, **32** (2009), 1316–1325. https://doi.org/10.1016/j.comcom.2009.02.002
6. N. Liu, M. Liu, W. Lou, G. Chen, J. Cao, PVA in VANETs: stopped cars are not silent, in *Proceedings of the 2011 IEEE International Conference on Computer Communications (INFOCOM)*, (2011), 431–435. https://doi.org/10.1109/INFCOM.2011.5935198
7. M. Liu, H. Gong, Y. Wen, G. Chen, J. Cao, The last minute: efficient data evacuation strategy for sensor networks in post-disaster applications, in *2011 Proceedings IEEE INFOCOM*, IEEE, Shanghai, China, (2011), 291–295. https://doi.org/10.1109/INFCOM.2011.5935131
8. N. Liu, M. Liu, G. Chen, J. Cao, The sharing at roadside: vehicular content distribution using parked vehicles, in *2012 Proceedings IEEE INFOCOM*, IEEE, Orlando, FL, USA, (2012), 2641–2645. https://doi.org/10.1109/INFCOM.2012.6195670
9. P. Wang, Q. Wu, C. Shen, A. van den Hengel, A. Dick, Explicit knowledge-based reasoning for visual question answering, preprint, arXiv:1511.02570.
10. P. Wang, Q. Wu, C. Shen, A. Dick, A. van den Hengel, Fvqa: fact-based visual question answering, *IEEE Trans. Pattern Anal. Mach. Intell.*, **40** (2017), 2413–2427. https://doi.org/10.1109/TPAMI.2017.2754246

11. M. Narasimhan, S. Lazebnik, A. Schwing, Out of the box: reasoning with graph convolution nets for factual visual question answering, in *Advances in Neural Information Processing Systems (NIPS)*, **31** (2018). Available from: https://papers.nips.cc/paper/2018/file/c26820b8a4c1b3c2aa868d6d57e14a79-Paper.pdf.
12. N. Kassner, H. Schütze, Negated and misprimed probes for pretrained language models: birds can talk, but cannot fly, preprint, arXiv:1911.03343.
13. H. Ren, W. Hu, J. Leskovec, Query2box: reasoning over knowledge graphs in vector space using box embeddings, preprint, arXiv:2002.05969.
14. H. Ren, J. Leskovec, Beta embeddings for multi-hop logical reasoning in knowledge graphs, in *Advances in Neural Information Processing Systems (NIPS)*, **33** (2020), 19716–19726.
15. B. Y. Lin, X. Chen, J. Chen, X. Ren, Kagnet: knowledge-aware graph networks for commonsense reasoning, preprint, arXiv:1909.02151.
16. A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in *Advances in Neural Information Processing Systems (NIPS)*, **26** (2013). Available from: https://papers.nips.cc/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf.
17. K. Guu, J. Miller, P. Liang, Traversing knowledge graphs in vector space, preprint, arXiv:1506.01094.
18. Y. Feng, X. Chen, B. Y. Lin, P. Wang, J. Yan, X. Ren, Scalable multi-hop relational reasoning for knowledge-aware question answering, preprint, arXiv:2005.00646.
19. M. Malinowski, M. Rohrbach, M. Fritz, Ask your neurons: a neural-based approach to answering questions about images, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (2015), 1–9. https://doi.org/10.1109/ICCV.2015.9
20. Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 21–29. https://doi.org/10.1109/CVPR.2016.10
21. D. Yu, J. Fu, T. Mei, Y. Rui, Multi-level attention networks for visual question answering, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 4709–4717. https://doi.org/10.1109/CVPR.2017.446
22. R. Hu, A. Rohrbach, T. Darrell, K. Saenko, Language-conditioned graph networks for relational reasoning, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 10294–10303. https://doi.org/10.1109/ICCV.2019.01039
23. P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao, A. van den Hengel, Neighbourhood watch: referring expression comprehension via language-guided graph attention networks, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 1960–1968. https://doi.org/10.1109/CVPR.2019.00206
24. L. Peng, S. Yang, Y. Bin, G. Wang, Progressive graph attention network for video question answering, in *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, (2021), 2871–2879. https://doi.org/10.1145/3474085.3475193

25. Y. Xi, Y. Zhang, S. Ding, S. Wan, Visual question answering model based on visual relationship detection, *Signal Process. Image Commun.*, **80** (2020), 115648. https://doi.org/10.1016/j.image.2019.115648
26. Y. Wu, Y. Ma, S. Wan, Multi-scale relation reasoning for multi-modal visual question answering, *Signal Process. Image Commun.*, **96** (2021), 116319. https://doi.org/10.1016/j.image.2021.116319
27. Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S. Y. Philip, A comprehensive survey on graph neural networks, *IEEE Trans. Neural Networks Learn. Syst.*, **32** (2020), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386
28. T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, preprint, arXiv:1609.02907.
29. M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, M. Welling, Modeling relational data with graph convolutional networks, in *Proceedings of the European Semantic Web Conference (ESWC)*, Springer, (2018), 593–607. https://doi.org/10.1007/978-3-319-93417-4_38
30. X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, et al., Heterogeneous graph attention network, in *The World Wide Web Conference*, (2019), 2022–2032. https://doi.org/10.1145/3308558.3313562
31. L. Hu, T. Yang, C. Shi, H. Ji, X. Li, Heterogeneous graph attention networks for semi-supervised short text classification, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (2019), 4821–4830. https://doi.org/10.18653/v1/D19-1488
32. K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD)*, (2008), 1247–1250. https://doi.org/10.1145/1376616.1376746
33. F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, et al., Language models as knowledge bases, preprint, arXiv:1909.01066.
34. A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, Y. Choi, Comet: commonsense transformers for knowledge graph construction, in *Association for Computational Linguistics (ACL)*, (2019), 4762–4779. https://doi.org/10.18653/v1/P19-1470
35. W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, et al., K-bert: enabling language representation with knowledge graph, in *Proceedings of the AAAI Conference on Artificial Intelligence (AI)*, **34** (2020), 2901–2908. https://doi.org/10.1609/aaai.v34i03.5681
36. J. Bao, N. Duan, Z. Yan, M. Zhou, T. Zhao, Constraint-based question answering with knowledge graph, in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (COLING)*, (2016), 2503–2514. Available from: https://aclanthology.org/C16-1236.pdf.
37. H. Sun, B. Dhingra, M. Zaheer, K. Mazaitis, R. Salakhutdinov, W. W. Cohen, Open domain question answering using early fusion of knowledge bases and text, preprint, arXiv:1809.00782.

38. P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, et al., Bottom-up and top-down attention for image captioning and visual question answering, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 6077–6086. Available from: https://openaccess.thecvf.com/content_cvpr_2018/papers/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.pdf.

39. M. Yasunaga, H. Ren, A. Bosselut, P. Liang, J. Leskovec, Qa-gnn: reasoning with language models and knowledge graphs for question answering, preprint, arXiv:2104.06378.

40. R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: an open multilingual graph of general knowledge, in *Thirty-first AAAI Conference on Artificial Intelligence (AI)*, **31**(2017). https://doi.org/10.1609/aaai.v31i1.11164

41. P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, preprint, arXiv:1710.10903.

42. T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al., Microsoft coco: common objects in context, in *European conference on computer vision (ECCV)*, Springer, (2014), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

43. Y. Zhu, O. Groth, M. Bernstein, F. F. Li, Visual7w: grounded question answering in images, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 4995–5004. https://doi.org/10.1109/CVPR.2016.540

44. R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, et al., Visual genome: connecting language and vision using crowdsourced dense image annotations, *Int. J. Comput. Vision*, **123** (2017), 32–73. https://doi.org/10.1007/s11263-016-0981-7

45. G. Li, H. Su, W. Zhu, Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks, preprint, arXiv:1712.00733.

AIMS Press