Knowledge and Information Systems (2024) 66:6975–7003
https://doi.org/10.1007/s10115-024-02166-8

REGULAR PAPER

# Caption matters: a new perspective for knowledge-based visual question answering

**Bin Feng[1] · Shulan Ruan[2] · Likang Wu[1] · Huijie Liu[1] · Kai Zhang[1] · Kun Zhang[3] · Qi Liu[1] · Enhong Chen[1]**

Received: 15 March 2024 / Revised: 10 June 2024 / Accepted: 16 June 2024 /
Published online: 22 July 2024


**Abstract**
Knowledge-based visual question answering (KB-VQA) requires to answer questions according to the given image with the assistance of external knowledge. Recently, researchers generally tend to design different multimodal networks to extract visual and text semantic features for KB-VQA. Despite the significant progress, 'caption' information, a textual form of image semantics, which can also provide visually non-obvious cues for the reasoning process, is often ignored. In this paper, we introduce a novel framework, the Knowledge Based Caption Enhanced Net (KBCEN), designed to integrate caption information into the KB-VQA process. Specifically, for better knowledge reasoning, we make utilization of caption information comprehensively from both explicit and implicit perspectives. For the former, we explicitly link caption entities to knowledge graph together with object tags and question entities. While for the latter, a pre-trained multimodal BERT with natural implicit knowledge is leveraged to co-represent caption tokens, object regions as well as question tokens. More-

✉ Bin Feng
fengbin1@mail.ustc.edu.cn

Shulan Ruan
slruan@mail.ustc.edu.cn

Likang Wu
wulk@mail.ustc.edu.cn

Huijie Liu
lhj33@mail.ustc.edu.cn

Kai Zhang
kkzhang08@ustc.edu.cn

Kun Zhang
zhang1028kun@gmail.com

Qi Liu
qiliuql@ustc.edu.cn

Enhong Chen
cheneh@ustc.edu.cn

1 State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China, Hefei 230026, China

2 Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China

3 School of Computer and Information, Hefei University of Technology, Hefei 230029, China

over, we develop a mutual correlation module to discern intricate correlations between explicit and implicit representations, thereby facilitating knowledge integration and final prediction. We conduct extensive experiments on three publicly available datasets (i.e., OK-VQA v1.0, OK-VQA v1.1 and A-OKVQA). Both quantitative and qualitative results demonstrate the superiority and rationality of our proposed KBCEN.

**Keywords** KB-VQA · Image captioning · Explicit and implicit learning · Knowledge reasoning · Mutual correlation

## 1 Introduction

Knowledge-based visual question answering (KB-VQA) [1] is a challenging task that requires an AI system to answer questions related to an image by leveraging external knowledge sources to enhance its prediction. KB-VQA harnesses three different sources of information: the input visual information (the image), the input textual information (the question) and external knowledge. By doing this, KB-VQA can help AI systems achieve a more human-like understanding of the world, which can enable practical applications such as improving image search and assisting visually impaired individuals. Although there has been great success in traditional VQA tasks [2–5], it is still full of challenges to achieve the human-like ability of co-comprehending images, questions and external knowledge together for KB-VQA. As a result, KB-VQA is an active area of research in natural language processing and computer vision, attracting increasing research interest in recent years.

With the successful accomplishments of various multimodal modeling methods, much progress has been made in this area. Most of those works focused on designing various networks to better extract image features, question features and link them with external knowledge. Early research [6] applied VGG [7] to extract image features, used Word2Vec to represent questions as word embeddings and constructed specific queries to retrieve supporting facts in the knowledge base. More recently, in order to better capture object regions and salient targets in the image, some works [8, 9] utilized Mask R-CNN [10], Faster R-CNN [11] as the visual feature extractor to extract a set of regional visual features. With the widespread application of vision-language pre-training (VLP) models, some VLP models, like ViLBERT [12] and LXMERT [13], improved visual representations for vision-language tasks from large-scale unsupervised datasets. In terms of introducing external knowledge for KB-VQA, previous studies designed various fusion networks for different types of knowledge, and there are some common approaches to utilizing external knowledge. The first common approach is parsing the knowledge in a symbolic format that usually consists of a collection of (subject, relation, object) triplets. For example, Gardères et al. [14] and Zhu et al. [15] designed graph structure networks to exploit ConceptNet KG for encoding structured knowledge. Different from that, another common approach is learning free-form knowledge in the network, which usually uses the raw contents as input. Among these networks, transformer-based architecture plays an extremely important role. For example, Lu et al. [12] and Tan et al. [13] built different multimodal transformer architectures, pre-trained on large-scale image-text datasets, to capture multimodal knowledge.

Although impressive progress has been accomplished with these efforts, there still exist several limitations to be unresolved. For example, the majority of the KB-VQA methods mentioned above only take into account visual features when utilizing images. However, in many instances, there is still a great deal of potentially useful information concealed in

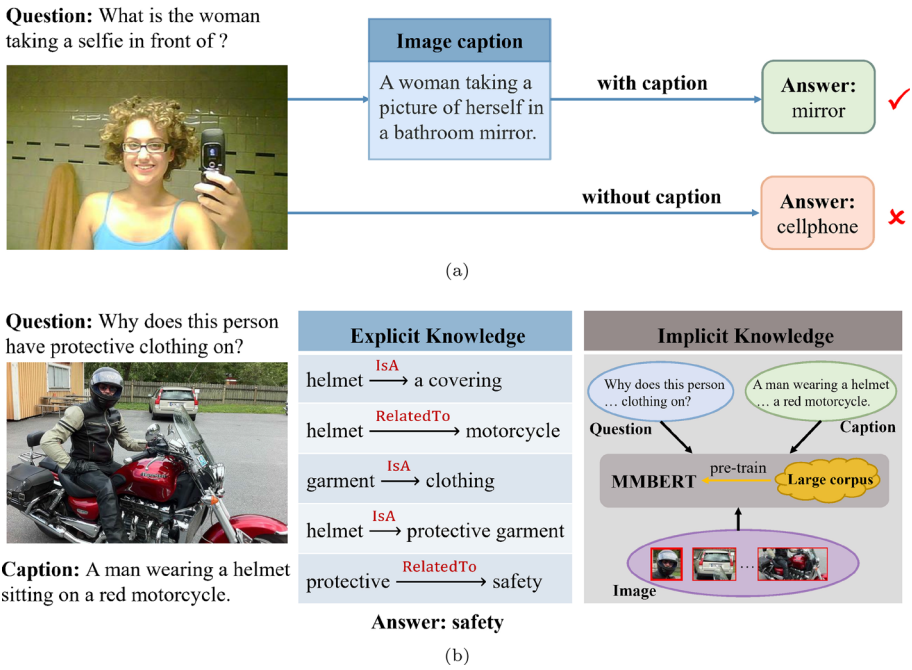



**Fig. 1** (a) is an example of the importance of image captions in knowledge reasoning. The model can correctly answer the question based on the image by using the caption. (b) is an example of the explicit and implicit knowledge contained in KB-VQA

the image, and relying solely on the analysis of visual features and the addition of pertinent external knowledge is insufficient to provide accurate answers. Taking Fig. 1a as an example, for a vision model (e.g., Faster R-CNN), it can easily recognize that there are *'woman'*, *'cellphone'* and *'towel'* in the image. Thus, the model might make a wrong prediction *'cellphone'* based on visually salient objects and simple common knowledge. However, when introducing caption information into the reasoning process, *'mirror'* then can be answered correctly. Although there is no mirror visually presented in the image, the caption can clearly describe that the woman needs to be in front of a bathroom mirror to take a selfie, no matter whether the target appears saliently or not in the image.

As a matter of fact, a caption is a textual form of image semantics, which can provide visually non-obvious cues for the reasoning process. Therefore, caption information is essential to improving image comprehension, especially for objects that are visually inconspicuous or invisible in the image. This naturally brings a new perspective for KB-VQA, which leads to the main focus of this paper: *how to make better utilization of caption information?* Unfortunately, inherent challenges persist in devising an effective way to incorporate caption information into KB-VQA. For one thing, the data in KB-VQA is multi-source and multimodal, including image data, text data, knowledge graph and so on. How to perform data processing and joint modeling on these structured and unstructured data is an important challenge to be resolved. For another, how to incorporate the caption into KB-VQA and make comprehensive utilization of it for better knowledge reasoning and integration is also an important challenge in this paper.

As illustrated in Fig. 1b, there usually exists rich external knowledge in KB-VQA tasks. Explicit knowledge (e.g., knowledge graph) usually originates from expert annotations, with a relatively small quantity but high quality. Implicit knowledge (e.g., pre-training) usually comes from task-specific pre-training on large-scale corpus with low interpretability. Therefore, different types of external knowledge can exhibit distinct data characteristics and coverage areas. This also to some extent inspires us to integrate caption information into the KB-VQA framework and jointly model multimodal data types in a comprehensive manner (i.e., explicitly and implicitly).

To this end, we propose a novel Knowledge Based Caption Enhanced Net (KBCEN) to leverage the caption and incorporate it into the reasoning and integration. Concretely, we first utilize Faster R-CNN [11] to extract visually salient object features and BERT [16] to learn question representations. Then, we make utilization of caption information comprehensively from both explicit and implicit perspectives for better reasoning. For the former, we explicitly link caption entities to knowledge graph (KG), together with object tags and question entities. While for the latter, a pre-trained multimodal BERT [17] with natural implicit knowledge is leveraged to co-represent caption tokens, object regions as well as question tokens. Finally, to model the unclear but complex correlations between explicit and implicit reasoning processes, we further develop a mutual correlation (MC) method for knowledge integration and final prediction.

To emphasize, we summarize the primary contributions of our work as follows:

- We observe the great potential of image caption for KB-VQA and propose to take both textual caption and visual information into consideration.
- We propose a novel KBCEN, in which a multimodal transformer and graph neural networks are designed to represent the multi-source and multimodal features comprehensively, i.e., explicitly and implicitly.
- To learn the complex but unclear correlations between explicit and implicit representations, we further develop a mutual correlation method to enhance each other from a mutual perspective.
- We conduct extensive experiments on three publicly available datasets (i.e., OK-VQA v1.0, OK-VQA v1.1 and A-OKVQA). Both quantitative and qualitative results demonstrate the superiority and rationality of our proposed KBCEN compared with baseline approaches.

The rest of the paper is organized as follows. In Sect. 2, we review literature that is closely related to our work. Next, our proposed method and technical details are presented in Sect. 3. Then, we conduct extensive experiments on benchmark datasets and perform detailed analyses in Sect. 4. Finally, we conclude our work and summarize possible future work in Sect. 5.

## 2 Related work

In this section, we will review the literature from three aspects: *Visual Question Answering*, *Knowledge-based Visual Question Answering*, as well as *Multimodal Vision and Language Modeling*, which are closely related to our work in this paper.

### 2.1 Visual question answering

Visual question answering (VQA), aiming to answer questions pertaining to a given image, has gained huge interest in recent years. In the earlier attempt at visual question answering, most studies of VQA are typically based on the CNN-RNN architecture [18–21]. For example, Malinowski et al. [18] integrated CNN and LSTM into an end-to-end architecture to predict answers conditioning on questions and corresponding images. For a better combination of image and question information, bilinear pooling methods [22, 23] have been proposed to fuse visual features from images with textual features from questions in fine-grained mode. Yu et al. [22] proposed a multimodal factorized bilinear pooling approach to fuse multimodal features. Ben-Younes et al. [23] leveraged a bilinear super-diagonal fusion strategy to optimize the trade-off between the expressiveness and complexity of the fusion mode. With the popularity of attention mechanisms and transformer architectures, various neural network modules [24–27] have been exploited in VQA tasks by adaptively learning the attended image features for a given question. Yang et al. [24] proposed a stacked attention network to learn the attention iteratively via multi-step reasoning. Anderson et al. [25] introduced a bottom-up and top-down attention mechanism at the level of objects and other salient regions. Liang et al. [26] designed a focal visual–text attention network for collective reasoning of visual and text sequence information. Changpinyo et al. [27] initialized the encoder from T5 and utilized a multilayer transformer to fuse the intermediate representation for textual question generation.

Despite the remarkable performance in traditional VQA tasks, they could only answer visual questions based on the given image and did not have a mechanism to incorporate the required knowledge from external sources. Reasoning approaches in the above work are always based on image and question features, which cannot be extended to involve external knowledge. To go one step further, in this paper, our study pays attention to not only original input features but also external knowledge during progressive reasoning.

### 2.2 Knowledge-based visual question answering

Based on the traditional VQA, the more challenging KB-VQA [1] was proposed, aiming to predict answers for general questions by leveraging external knowledge beyond image content. Multiple KB-VQA datasets have been proposed and played a crucial role in KB-VQA research [6, 28–30]. In [28], the early KB-VQA dataset only involved 700 images from the MSCOCO validation set and 2402 questions generated by predefined templates. FVQA dataset [6] contained 2190 images and 5826 questions, in which each question–answer sample was annotated with its own knowledge bases and ground-truth supporting facts. The dataset provided a supporting fact for '*question–answer*' pairs in the form of a structural triplet (*image, question, answer, supporting fact*). OK-VQA v1.0 dataset [1] was the first large-scale dataset with questions that needed to be answered using open-world knowledge instead of a provided fixed knowledge base. OK-VQA v1.1 dataset [9] was a more recent open-domain dataset that covered a wide range of topics and included 14,055 questions based on 14,031 images. A-OKVQA dataset [29], based on OK-VQA datasets, was specifically targeted for KB-VQA on open-domain natural scenes.

Recent methods for KB-VQA mainly focused on designing various networks to introduce various knowledge to solve KB-VQA tasks, including knowledge graph-based approaches [14, 15], unstructured knowledge-based approaches [31, 32], implicit knowledge-based pre-training approaches [12, 13] and multi-source knowledge-based hybrid approaches [8, 9]. For

example, Gardères et al. [14] exploited and evaluated the knowledge graph ConceptNet for encoding common sense knowledge on the OK-VQA dataset. Gao et al. [31] used Wikipedia passages as the external knowledge base and retrieved top-k relevant knowledge for the KB-VQA task. Lu et al. [12] pre-trained on large-scale image-text datasets and then fine-tuned it on the specific KB-VQA task. Wu et al. [8] constructed a multi-source hybrid knowledge pool, including Wikipedia, ConceptNet and Google Images, to query the most relevant knowledge. Motivated by the promising capacity of LLMs, some works [33–38], based on massive model scale and computing power, achieved comparable results with other models trained on the specific KB-VQA dataset. Lin et al. [33] transformed the image into plain text and encoded the question, visual context and knowledge passages into T5 model for answer generation. Prophet [39], Promptcap [40] and KAT [41], based on GPT-3 with 175 Billion parameters, achieved good performance. Shao et al. [34] introduced answer heuristics generation and heuristics-enhanced prompting to activate the capacity of GPT-3. Lin et al. [35] prompted GPT-3 by regional tags, question and context in natural language space to retrieve external knowledge.

However, most of the aforementioned methods did not fully explore the relationships and correlations between different types of knowledge effectively. Explicit knowledge (e.g., knowledge graph) usually originates from expert annotations, with a relatively small quantity but high quality. Implicit knowledge (e.g., pre-training) usually comes from task-specific pre-training on large-scale corpus with low interpretability. Therefore, different types of external knowledge can exhibit distinct data characteristics and coverage areas. Jointly learning from both explicit and implicit knowledge can be helpful for knowledge integration and answer prediction.

### 2.3 Multimodal vision and language modeling

With the rapid development of vision-language modeling techniques, research on vision and language tasks has explored a wide range of multimodal fusion strategies [42–46]. In the earlier attempt at vision-language modeling, simple concatenation or element-wise multiplication between visual and linguistic elements was proposed for cross-modal feature fusion [47]. To capture high-level interactions between multiple modalities, Fukui et al. [48] proposed to utilize multimodal compact bilinear pooling to project image and text representations to a higher dimensional space. Gao et al. [49] combined cross-modal self-attention and co-attention mechanisms to ensure efficient information exchange within and across image and language modalities. Hannan et al. [50] considered three distinct modalities (text, images and tables) and explored the interactions between natural language and other modalities.

With the great success of Transformer [51] and BERT [16] in natural language processing fields, many recent multimodal works have been inspired and proposed transformer-based fusion of multiple modalities [52–57]. Based on the model structures, these transformer-based methods can be broadly classified into two types: single-stream and two-stream. VisualBERT [17], Unicoder-VL [58] and VL-BERT [59] proposed the single-stream architecture to work on both images and text simultaneously. ViLBERT [12] and LXMERT [13] proposed the two-stream framework to deal with images and text separately and fuse them by another transformer in the next phase. With the exploration of a unified architecture for multimodality, more recent works (e.g., CLIP [60], BLIP [61], ImageBind [62]) have been successively proposed. These methods have demonstrated their capacity to retain comprehensive cross-modal understanding, resulting in remarkable performance in various multimodal tasks.

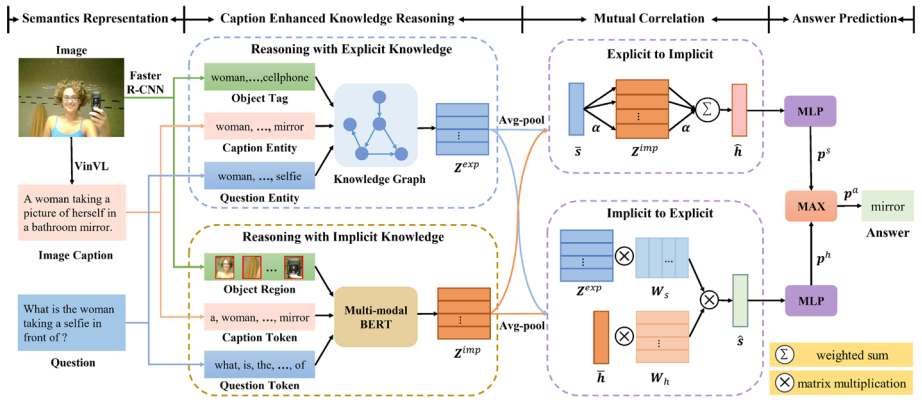



**Fig. 2** The overall architecture of Knowledge Based Caption Enhanced Net (KBCEN) for knowledge-based visual question answering

The development of these methods has greatly promoted the advancement of KB-VQA, one of the essences of which is also to better jointly model images and text. However, the great potential of caption information is often ignored in prior research. It can provide visually non-obvious cues for the reasoning process. Therefore, in this paper, we advocate for greater emphasis on caption information and introduce a novel Knowledge Based Caption Enhanced Net (KBCEN) to make utilization of caption information comprehensively from both explicit and implicit perspectives.

## 3 Method

In this section, we mainly introduce the problem statement and technical details of our proposed Knowledge Based Caption Enhanced Net (KBCEN).

### 3.1 Problem statement

First of all, we formally define our task here. Given an image $\boldsymbol{I}$, its caption $\boldsymbol{C}$, a question $\boldsymbol{Q}$ related to the image, external knowledge $\boldsymbol{K}$ as well as an answer vocabulary set $\boldsymbol{A}$, our goal is to learn a KB-VQA model $\xi$ which is able to precisely predict the meaningful answer $\boldsymbol{a} \in \boldsymbol{A}$, with $\boldsymbol{a} = \xi(\boldsymbol{I}, \boldsymbol{C}, \boldsymbol{Q}, \boldsymbol{K})$.

Notably, the statement here is actually a general form of problem definition. If $\boldsymbol{C}$ is available in the dataset, it could be directly used as the task input. Otherwise, we can generate it in our framework with pre-trained image captioning models (e.g., VinVL [63]). Both cases are applicable to our proposed method.

To achieve this goal, as depicted in Fig. 2, we propose a novel KBCEN which embodies four main components: (1) *Semantics Representation*: introducing captions, extracting visual and question features; (2) *Caption Enhanced Knowledge Reasoning*: reasoning with implicit and explicit knowledge; (3) *Mutual Correlation*: mutual correlating two different feature representations for information enhancement; (4) *Answer Prediction*: predicting the answer robustly. The details will be introduced in the following parts.

## 3.2 Semantics representation

### 3.2.1 Image representation

Since both visual features and caption semantics from the image are vital for knowledge-based visual question answering, we extract them simultaneously in this component.

Faster R-CNN [11] has powerful capabilities for image representation and salient object detection. Therefore, in this paper, following [14, 15], we employ Faster R-CNN, pre-trained on Visual Genome [64], as the visual feature extractor. Specifically, we apply Faster R-CNN to detect a set of object regions $\boldsymbol{O} = \{\boldsymbol{o}_i\}_{i=1}^{n_o}$ and corresponding object tags $\boldsymbol{T} = \{\boldsymbol{w}_i\}_{i=1}^{n_t}$ in the image. We represent each object $\boldsymbol{o}_i$ by a visual feature vector $\boldsymbol{f}_i \in \mathbb{R}^{d_f}$ ($d_f = 2048$) and a spatial feature vector $\boldsymbol{b}_i \in \mathbb{R}^{d_b}$ ($d_b = 4$):

$$\boldsymbol{O}, \boldsymbol{T} = f_{OD}(\boldsymbol{I}), \tag{1}$$

where $f_{OD}(\cdot)$ is object detection model, e.g., Faster R-CNN.

A comprehensive understanding of image content plays a vital role in visual question answering. Apart from object-centric visual features, image-level caption information has been proven to be crucial in image content understanding, as demonstrated in [65, 66]. When humans are asked about the content of an image, they usually recognize the salient objects, consolidate their relationships, and then apply common sense knowledge to describe them syntactically and semantically. Inspired by human behavior, we introduce image-level caption information, since image caption mimics the remarkable human ability by compressing a set of salient visual information into descriptive language.

As mentioned above, the caption can provide visually non-obvious or invisible cues for reasoning, such as Fig. 1a. Therefore, to comprehensively represent the image, in this paper, we view caption information $\boldsymbol{C}$ as a textual form of image semantics. In particular, if the caption is not provided in the dataset, such as those VQA-related datasets, we can employ an image captioning model to generate it. In order to better understand the semantic information of an image and generate its accurate description, as shown in Fig. 2, we select a state-of-the-art image captioning model, specifically VinVL [63], to transform image-level information to caption text.

### 3.2.2 Text representation

BERT [16] has been proven to be a great success in various natural language processing tasks, especially for text encoding. Therefore, following [8, 14], we utilize BERT as the text encoder in this paper. As shown in Fig. 2, we employ BERT to process and represent both question (i.e., $\boldsymbol{Q}$) and the generated image caption (i.e., $\boldsymbol{C}$) in this paper. To be specific, we tokenize $\boldsymbol{Q}$ and $\boldsymbol{C}$ using WordPiece [67] as in BERT to generate sequences of discrete tokens, consisting of vocabulary words and a small set of special tokens, i.e., '`SEP`,' '`CLS`,' '`MASK`.' Then, we embed them with pre-trained BERT embeddings to generate sequences of $d_s$-dimensional token representation $\boldsymbol{Q} = \{\boldsymbol{w}_1^q, ..., \boldsymbol{w}_{n_q}^q\} \in \mathbb{R}^{n_q \times d_s}$ and $\boldsymbol{C} = \{\boldsymbol{w}_1^c, ..., \boldsymbol{w}_{n_c}^c\} \in \mathbb{R}^{n_c \times d_s}$. The representation of each token is composed of three parts: a token-specific learned embedding, an encoding for its position within the sequence and an encoding for its segment, which shows the index of the token's sentence if multiple sentences exist.

### 3.3 Caption enhanced knowledge reasoning

In the Semantics Representation component, we have obtained question features and represented visual features comprehensively from multiple aspects (e.g., salient objects, object tags and image caption). How to better leverage image caption and incorporate it into knowledge reasoning is still very challenging. To this end, we propose to make utilization of caption information comprehensively from both explicit and implicit perspectives.

#### 3.3.1 Reasoning with explicit knowledge

To better answer visual questions in KB-VQA, it is necessary to introduce external knowledge from various knowledge sources. Unlike previous work such as FVQA [6], we do not have a ground-truth set of facts or knowledge which can be used to answer the question. We first need to choose which knowledge sources to use and how to integrate and represent them.

Following [9, 14], we choose four different knowledge sources to construct our knowledge graphs, i.e., ConceptNet [68], DBPedia [69], VisualGenome [64] and hasPartKB [70]. They cover crowdsourced data, encyclopedic data, visual data, knowledge about everyday objects, science, specific people, places and events. For better knowledge representation, we use the relational graph convolutional network (RGCN) [71] as the base graph network for our model. Unlike the related GCN [72], RGCN uses different weight matrices for different edge types (the *'is_a'* relationship is different from the *'has_a'* relationship) and different edge directions (*'vegetable is_a food'* is different from *'food is_a vegetable'*). Therefore, the differences in edge types and edge directions can be easily captured and well represented by this type of network structure.

With these knowledge sources, we can capture a vast amount of knowledge about the world, and more details of these knowledge graphs are described as follows.

ConceptNet is a knowledge graph built from several different primary sources, including Wiktionary, Open Mind Common Sense and Games with a purpose. It contains over 21 million edges and over 8 million nodes, representing words and phrases widely used and common sense relationships between them. DBPedia is a knowledge graph collected from Wikipedia articles and tables. It allows you to ask sophisticated queries against datasets derived from Wikipedia and to link other datasets on the Web to Wikipedia data. VisualGenome is a knowledge graph built from the intersection of the YFCC100M and MSCOCO. It can represent the interactions and relationships between objects in an image and include objects, attributes, relationships and noun phrases in region descriptions. hasPartKB is a knowledge graph with *'hasPart'* relationships extracted from a large corpus of generic statements. It contains 49,848 edges with information about quantifiers, argument modifiers and links the entities to appropriate concepts in Wikipedia and WordNet.

To integrate different knowledge sources, we first fuse all knowledge triplets from the four knowledge graphs. Then, remove all stop words (e.g., '`is`,' '`the`,' '`a`') from the set to avoid non-meaningful edges. Next, we collect all of the semantic concepts from the dataset and then include edges that only include these concepts. After this filtering, we generate a final knowledge subgraph. According to the statistics of the experimental results, the final graph includes 379,624 edges and 8471 nodes, and an average of 24.65 nodes is activated per question.

To better leverage explicit knowledge and align with graph nodes, we extract all of the semantic concepts from the input (i.e., object tags, caption entities and question entities). In this process, we utilize BERT to segment the caption and question text to obtain candidate

entities. Meanwhile, we also use the visual recognition system (i.e., Faster R-CNN) to generate and pick up object tags. Then, RGCN receives the input denoted as the $\boldsymbol{H}^{(0)} \in \mathbb{R}^{n \times d_s}$ layer with $n$ node inputs of size $d_s$. For each layer in RGCN, we have a nonlinear function $\boldsymbol{H}^{(l+1)} = f(\boldsymbol{H}^{(l)}, \boldsymbol{K})$ where $\boldsymbol{K}$ is the knowledge graph. After all RGCN layers are computed, we end up with the explicit embedding matrix $\boldsymbol{Z}^{exp} \in \mathbb{R}^{d_e \times d_s}$.

#### 3.3.2 Reasoning with implicit knowledge

Based on [8, 9], we further extend the multimodal BERT to better incorporate caption information into the implicit reasoning process. The extended multimodal BERT is pre-trained on BooksCorpus (800 M words) and English Wikipedia (2.5B words), which naturally introduces implicit knowledge into KB-VQA. Since vision-language pre-training has shown great potential for many multimodal tasks, we further pre-train the multimodal BERT on the training set of KB-VQA-related dataset (i.e., VQAv2 [73]). As depicted in Fig. 2, we build the multimodal transformer model and input object regions, caption tokens as well as question tokens into the extended multimodal BERT together for joint learning and reasoning to generate the implicit embedding matrix $\boldsymbol{Z}^{imp} \in \mathbb{R}^{d_m \times d_h}$.

### 3.4 Mutual correlation

In the Caption Enhanced Knowledge Reasoning component, joint representations from two perspectives (i.e., with explicit or implicit knowledge) have been obtained. However, the two features should not be independent of each other, since they have similar inputs and the same learning goal. Therefore, there might be some complex but unclear correlation between them. How to model and utilize the correlation is crucial for better joint representation and performance improvement. As shown in Fig. 2, we develop a novel mutual correlation (MC) module to learn the unclear but complex relation between explicit and implicit knowledge-enhanced representations. The technical details are introduced as follows.

#### 3.4.1 Implicit to explicit

As mentioned above, incorporating implicit knowledge plays a vital role in judging the correct answer in the KB-VQA task. It seems viable to leverage implicit knowledge to enrich explicit representation. Specifically, as shown in Fig. 2, we first transform implicit embedding matrix $\boldsymbol{Z}^{imp}$ into implicit feature vector $\bar{\boldsymbol{h}}$ with average pooling. Then, we leverage matrix multiplication to generate the well-learned explicit representation $\hat{\boldsymbol{s}}$. The implicit representation is reused for enhancing explicit embedding, and this gives us a late fusion between the implicit and explicit parts in the mutual correlation module. The process is formulated as follows:

$$\begin{aligned} \bar{\boldsymbol{h}} &= avg\_pooling(\boldsymbol{Z}^{imp}), \\ \hat{\boldsymbol{s}}_i &= \sigma((\boldsymbol{W}_s \boldsymbol{Z}_i^{exp} + \boldsymbol{b}_s)^T (\boldsymbol{W}_h \bar{\boldsymbol{h}} + \boldsymbol{b}_h)), \end{aligned} \tag{2}$$

where $\boldsymbol{W}_s$, $\boldsymbol{W}_h$ are trainable parameters, and $avg\_pooling(\cdot)$ means average pooling.

#### 3.4.2 Explicit to implicit

Since explicit representation from KGs covers the wide variety of knowledge needed to answer visual questions, it is also vital to enhance the implicit representation with the explicit

one. As is known to all, attention mechanism plays a crucial role in extracting the most relevant parts from inputs for outputs [74, 75]. Therefore, as shown in Fig. 2, with the query of the explicit feature, we adopt attention mechanism to calculate different weights of vectors in implicit feature matrix (i.e., $\boldsymbol{Z}^{imp}$). It could be computed as follows:

$$\begin{aligned} \bar{\boldsymbol{s}} &= avg\_pooling(\boldsymbol{Z}^{exp}), \\ \boldsymbol{M}_i &= tanh(\boldsymbol{W}_z \boldsymbol{Z}_i^{imp} + \boldsymbol{W}_p \bar{\boldsymbol{s}}), \\ \boldsymbol{\alpha} &= softmax(\boldsymbol{W}_a^T \boldsymbol{M}), \\ \hat{\boldsymbol{h}} &= \sum_{i=1}^{d_m} \alpha_i \boldsymbol{Z}_i^{imp}, \end{aligned} \tag{3}$$

where $\boldsymbol{W}_z$, $\boldsymbol{W}_p$, $\boldsymbol{W}_a$ are trainable parameters. $\boldsymbol{\alpha}$ denotes the attention weight computed from the avg-pooled explicit vector $\bar{\boldsymbol{s}}$ to implicit feature matrix $\boldsymbol{Z}^{imp}$. Then, $\boldsymbol{\alpha}_i$ is utilized to calculate a weighted sum (i.e., $\hat{\boldsymbol{h}}$) of different vectors in the implicit feature matrix.

### 3.5 Answer prediction and model learning

#### 3.5.1 Answer prediction

With our proposed MC, we could obtain well-learned multimodal representation vectors (i.e., $\hat{\boldsymbol{h}}$ and $\hat{\boldsymbol{s}}$). Next, we employ two multilayer perceptrons (MLPs) to calculate the answer separately. Specifically, each MLP consists of two hidden layers with $ReLu$ activation function and a $softmax$ output layer, which can be formulated as follows:

$$\boldsymbol{p}^s = \mathrm{MLP}_1(\hat{\boldsymbol{s}}),\ \boldsymbol{p}^h = \mathrm{MLP}_2(\hat{\boldsymbol{h}}), \tag{4}$$

where $\boldsymbol{p}^s$ and $\boldsymbol{p}^h$ indicate the probability distribution vectors of explicit and implicit representations about candidate answers, respectively. Finally, we make the final prediction by choosing the highest scoring answer from the two vectors.

$$\boldsymbol{p}^a = max(\boldsymbol{p}^s, \boldsymbol{p}^h). \tag{5}$$

#### 3.5.2 Model learning

Since KB-VQA is formulated as a classification problem in this paper, we adopt cross-entropy as the loss function:

$$L = -\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i log P(\boldsymbol{p}_i^a \mid \boldsymbol{I}, \boldsymbol{Q}, \boldsymbol{C}), \tag{6}$$

where $\boldsymbol{y}_i$ is the true answer label of the $i^{th}$ instance of the dataset, and $n$ represents the number of training instances.

## 4 Experiment

In this section, we first introduce the experiment preparation, involving the data description and experiment setting. Then, we evaluate the model performance on three public benchmark

**Table 1** Statistics of OK-VQA v1.0, v1.1 and A-OKVQA datasets for knowledge-based VQA. Specifically, we show the number of images, number of questions, average question length, average answer length and answer processing method in each dataset

| Dataset | #Image | #Question | Question length | Answer length | Answer processing |
|---|---|---|---|---|---|
| OK-VQA v1.0 [1] | 14,031 | 14,055 | 8.1 | 1.3 | Raw |
| OK-VQA v1.1 [9] | 14,031 | 14,055 | 8.1 | 1.3 | Word stemming |
| A-OKVQA [29] | 23,692 | 24,903 | 8.8 | 1.3 | Word stemming |

datasets and conduct some ablation studies. Next, we give a detailed analysis of the model and experiment results. Moreover, we also present both quantitative and qualitative studies to demonstrate the superiority and rationality of our proposed method.

### 4.1 Data description

We evaluate our method on Outside Knowledge Visual Question Answering Dataset (OK-VQA) v1.0 [1], OK-VQA v1.1 [9], as well as Augmented successor of OK-VQA (A-OKVQA) [29]. The three datasets are specifically targeted for the knowledge-based VQA task on open-domain natural scenes, only including questions that require external resources to answer.

OK-VQA v1.0 is composed of 14,031 images and 14,055 questions. Its questions cover a variety of 10 knowledge categories, and are annotated by Amazon Mechanical Turkers. Each data sample is made up of one image, one corresponding question and 10 ground-truth answers. The training and testing sets consist of 9009 and 5046 samples, respectively.

Compared with the v1.0 version, the OK-VQA v1.1 dataset uses a good word-stemming method on the raw answers to improve their quality, resulting in a more coherent answer vocabulary. They have the same data scale and split.

A-OKVQA contains a diverse set of 23,692 images and 24,903 questions requiring a broad base of commonsense and world knowledge to answer. Questions in A-OKVQA are challenging, conceptually diverse and require knowledge outside the image. Compared to existing knowledge-based visual question answering datasets, they cannot be answered by simply querying the knowledge base.

To show these datasets more clearly and intuitively, we perform statistical analysis of these datasets for the knowledge-based VQA task, shown in Table 1. We also show the distribution of knowledge types required for answering questions on different datasets in Fig. 3.

### 4.2 Experiment setting

#### 4.2.1 Model setting

For multimodal BERT, we initialize the network from BERT-base with 12 layers, a hidden size of 768 and 12 attention heads.

For the knowledge graph module, we initialize the base graph network from RGCN with two convolutional layers, a node hidden size of 128. The max sentence length of the image caption and the question is set to 128. The visual embedding dimension of the image is set to 2048. For explicit and implicit embedding matrix dimensions, we set $(d_e, d_s, d_m, d_h) = (1746, 128, 228, 768)$.

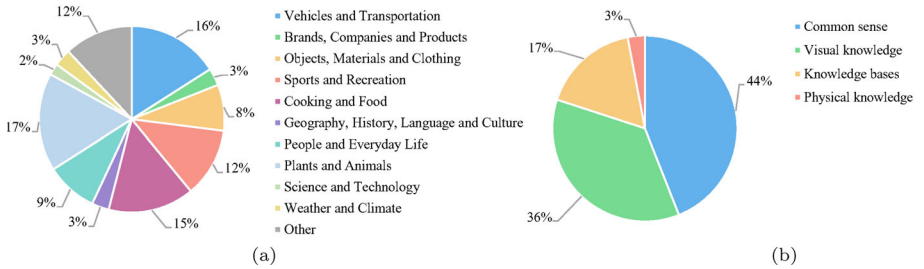



**Fig. 3** The distribution of knowledge types required for answering questions on different datasets. **a** We show the percentage of visual questions falling into 10 knowledge categories on OK-VQA v1.0 and v1.1 datasets. **b** We show the knowledge type distribution on the A-OKVQA dataset

#### 4.2.2 Training setting

To initialize the model, we set all weights such as $\boldsymbol{W}$ following the truncated normal distribution, and use AdamW optimizer with the learning rate of $5 \times 10^{-5}$, warmup scheduler of cosine, warmup steps of 2000. Furthermore, we set the batch size to 32 per GPU and the total training epochs to 120. Our model is implemented with *PyTorch* and trained on 2 Nvidia Tesla V100 GPUs.

#### 4.2.3 Evaluation metrics

Following [76, 77], for better comparison, we also adopt the standard VQA evaluation metric to measure the performance of different methods. Each question has ten ground-truth answers annotated by ten different people, where people who provide answers are not the same as people who ask questions. The answers are evaluated with the standard VQA accuracy metric as follows:

$$\text{Acc(ans)} = \min\left(1, \frac{\#\{\text{humans provided ans}\}}{3}\right), \tag{7}$$

where #{humans provided ans} means the number of humans that provided the answer. In addition, the top-1 accuracy and top-3 accuracy are calculated for each method.

Our project codes as well as data are all publicly available at https://github.com/zzmyrep/KBCEN.

### 4.3 Baselines

In this paper, we compare our model against the following baselines:

- *MLP* [1]: having three hidden layers with ReLU activation and a hidden size of 2048 with the image and question features as input.
- *BAN* [2]: leveraging a co-attention mechanism between the bottom-up detection features of the image and the question features.
- *BAN+KG-Aug* [3]: incorporating the aggregated external knowledge from knowledge graphs into BAN through a context-aware fusion mechanism that requires no ground-truth facts as supervised guidelines.

- *MUTAN* [4]: utilizing a multimodal tensor-based Tucker decomposition, which encodes full second-order interactions, to efficiently parametrize bilinear interactions between visual and textual modalities.
- *Mucko* [15]: proposing a modality-aware heterogeneous graph network to capture complementary evidence that is most close to the visual question.
- *ViLBERT* [12]: extending the BERT architecture to a multimodal two-stream model, processing both visual and textual inputs in separate streams that interact through co-attentional transformer layers.
- *LXMERT* [13]: consisting of a natural language encoder, an object relationship encoder and a cross-modality encoder, and then pre-trained with diverse representative pre-training tasks.
- *ConceptBERT* [14]: exploiting transformer blocks and knowledge graphs to enhance the representation and then aggregating multiple embeddings to learn a joint concept-vision-language embedding.
- *CBM*$_{\mathbf{BERT}}$ [54]: utilizing OSCAR to generate captions of images for verbalizing the visual contents and applying a pre-trained text-only language model such as BERT for inference.
- *KRISP* [9]: exploiting transformer-based network for vision-language alignment and integrating symbolic representations from diverse knowledge graphs to solve knowledge-based questions.
- *MAVEx* [8]: proposing an answer validation method by extracting relevant knowledge from noisy sources and verifying the validity of each candidate based on the retrieved knowledge.
- *PGVQA* [78]: estimating the possibility of candidate answers based on the purpose of the question to be answered, and incorporating knowledge facts that match the question purpose into answer prediction.
- *MuKEA* [53]: employing a heteroid triplet to establish correlations between visual images and factual answers. For the purpose of answering predictions, both fundamental and domain-specific information are progressively accumulated.
- *MSG-KRM* [55]: bridging the multimodal semantic gap by embedding heterogeneous graph and data as feature vectors into the same dimension, and incorporating type-aware feature vectors into semantic graphs to enhance the inference capability.

### 4.4 Experiment result

We compare our model against the following baselines: MLP [1], BAN [2], BAN+KG-Aug [3], MUTAN [4], Mucko [15], ViLBERT [12], LXMERT [13], ConceptBERT [14], CBM$_{\text{BERT}}$ [54], KRISP [9], MAVEx [8], PGVQA [78], MuKEA [53] and MSG-KRM [55]. In addition, for better comparison, we also conduct experiments on these baselines that take the text caption into account, instead of only using the question and the corresponding image. The suffix '-Cap' means that we enhance baseline methods by including the caption as input in addition to the question and the corresponding image. The overall results on three benchmark datasets are summarized in Table 2, which shows our model achieves the best performance.

Experiments (1)–(8) are traditional methods dealing with VQA. By comparison, we can observe that these traditional methods fail to obtain good solutions to the KB-VQA. They cannot obtain deep semantic features of images and draw upon outside knowledge to answer questions, which leads to very limited performance for KB-VQA.

**Table 2** Overall performance (accuracy) of different methods on OK-VQA v1.0, v1.1 and A-OKVQA datasets

| Model | Approx. Params (M) | OK-VQA v1.0 | | OK-VQA v1.1 | | A-OKVQA | |
|---|---|---|---|---|---|---|---|
| | | Top-1 (%) | Top-3 (%) | Top-1 (%) | Top-3 (%) | Top-1 (%) | Top-3 (%) |
| (1) MLP [1] | 8 | 20.67 | 23.09 | 20.84 | 23.64 | 17.32 | 20.33 |
| (2) MLP-Cap [1] | 8 | 21.43 | 24.11 | 21.68 | 24.78 | 18.23 | 21.47 |
| (3) BAN [2] | 45 | 25.17 | 27.76 | 25.42 | 28.10 | 22.75 | 24.51 |
| (4) BAN-Cap [2] | 45 | 26.24 | 28.92 | 26.57 | 29.58 | 24.17 | 25.79 |
| (5) BAN+KG-Aug [3] | 51 | 26.71 | 29.64 | 26.85 | 30.07 | 23.92 | 25.07 |
| (6) BAN+KG-Aug-Cap [3] | 51 | 28.34 | 31.29 | 28.49 | 31.78 | 24.75 | 26.63 |
| (7) MUTAN [4] | 70 | 26.41 | 29.32 | 26.79 | 29.84 | 24.10 | 25.36 |
| (8) MUTAN-Cap [4] | 70 | 28.16 | 30.81 | 28.23 | 31.61 | 25.23 | 26.77 |
| (9) Mucko [15] | 118 | 29.20 | 30.66 | 29.61 | 32.39 | 27.32 | 29.83 |
| (10) Mucko-Cap [15] | 118 | 30.98 | 32.83 | 31.59 | 34.47 | 28.42 | 31.28 |
| (11) ViLBERT [12] | 116 | 31.35 | 34.69 | 31.81 | 34.80 | 30.60 | 33.64 |
| (12) ViLBERT-Cap [12] | 116 | 33.87 | 38.05 | 34.63 | 38.52 | 32.19 | 35.27 |
| (13) LXMERT [13] | 340 | 32.04 | 36.85 | 32.49 | 37.11 | 30.70 | 33.89 |
| (14) LXMERT-Cap [13] | 340 | 34.73 | 39.59 | 35.16 | 39.84 | 32.43 | 35.83 |
| (15) ConceptBERT [14] | 118 | 33.66 | 37.42 | 33.87 | 38.05 | 31.87 | 35.17 |
| (16) ConceptBERT-Cap [14] | 118 | 36.82 | 41.17 | 37.54 | 42.29 | 33.41 | 37.43 |
| (17) $CBM_{BERT}$ [54] | 112 | 36.00 | 39.25 | 36.73 | 40.11 | 33.49 | 36.86 |
| (18) $CBM_{BERT}$-Cap [54] | 112 | 37.34 | 42.16 | 38.27 | 42.95 | 34.18 | 38.27 |
| (19) KRISP [9] | 116 | 38.35 | 42.68 | 38.90 | 43.18 | 33.71 | 38.02 |
| (20) KRISP-Cap [9] | 116 | 41.03 | 45.80 | 41.67 | 46.93 | 35.63 | 40.58 |
| (21) MAVEx [8] | 353 | 39.38 | 44.56 | 39.45 | 45.34 | 34.10 | 40.25 |
| (22) MAVEx-Cap [8] | 353 | 40.67 | 45.89 | 41.09 | 46.52 | 35.81 | 41.73 |

**Table 2** continued

| Model | Approx. Params (M) | OK-VQA v1.0 | | OK-VQA v1.1 | | A-OKVQA | |
|---|---|---|---|---|---|---|---|
| | | Top-1 (%) | Top-3 (%) | Top-1 (%) | Top-3 (%) | Top-1 (%) | Top-3 (%) |
| (23) PGVQA [78] | 120 | 41.07 | 46.13 | 40.36 | 46.08 | 35.97 | 41.96 |
| (24) PGVQA-Cap [78] | 120 | 41.84 | 46.95 | 42.16 | 47.65 | 37.04 | 43.05 |
| (25) MuKEA [53] | 342 | 42.02 | 47.34 | 42.59 | 48.60 | 38.17 | 43.49 |
| (26) MuKEA-Cap [53] | 342 | 42.96 | 48.87 | 43.15 | 50.08 | 40.23 | 45.18 |
| (27) MSG-KRM [55] | 450 | 43.12 | 48.53 | 43.58 | 49.26 | 38.78 | 44.35 |
| (28) MSG-KRM-Cap [55] | 450 | 43.67 | 49.14 | 43.74 | 50.63 | 41.02 | 46.24 |
| (29) KBCEN | 116 | 45.73 | 52.08 | 46.30 | 53.91 | 44.78 | 50.52 |

**Table 3** Ablation study of KBCEN on OK-VQA v1.0, v1.1 and A-OKVQA datasets, where w/o means without

| Model | OK-VQA v1.0 | | OK-VQA v1.1 | | A-OKVQA | |
|---|---|---|---|---|---|---|
| | Top-1 (%) | Top-3 (%) | Top-1 (%) | Top-3 (%) | Top-1 (%) | Top-3 (%) |
| (1) KBCEN(w/o caption) | 39.32 | 45.73 | 40.45 | 48.24 | 38.82 | 44.79 |
| (2) KBCEN(w/o KG) | 42.84 | 49.03 | 43.49 | 50.77 | 41.97 | 47.58 |
| (3) KBCEN(w/o MMBERT) | 40.97 | 47.46 | 41.25 | 49.01 | 40.08 | 46.03 |
| (4) KBCEN(w/o multiply in MC) | 44.80 | 50.74 | 45.26 | 52.85 | 43.93 | 49.79 |
| (5) KBCEN(w/o attention in MC) | 45.08 | 51.21 | 45.47 | 53.34 | 43.16 | 48.95 |
| (6) KBCEN(w/o MC) | 44.52 | 50.75 | 44.95 | 52.26 | 42.50 | 47.68 |
| (7) KBCEN | 45.73 | 52.08 | 46.30 | 53.91 | 44.78 | 50.52 |

Experiments (9)–(28) are recent KB-VQA methods. For VLP-based models like ViLBERT and LXMERT, they incorporate implicit knowledge by pre-training on the large-scale corpus, but ignore combining explicit knowledge from external high-quality resources. For KG-based approaches like Mucko and ConceptBERT, they leverage ConceptNet to retrieve the supporting knowledge and conduct graph reasoning for answer prediction. However, they do not utilize knowledge comprehensively. $\text{CBM}_{\text{BERT}}$ defines a text-only method to take advantage of implicit knowledge of LMs, but underexplores explicit knowledge and visual features. KRISP, MAVEx, PGVQA, MuKEA and MSG-KRM methods leverage different types of knowledge, but overlook the semantic information of image captions, which require a comprehensive understanding of the image.

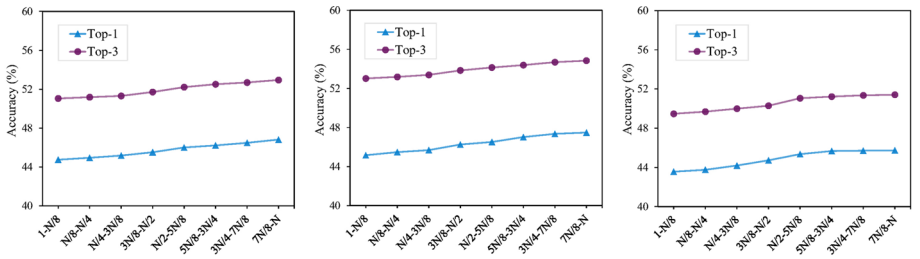
Our proposed KBCEN introduces the caption information into KB-VQA, which could provide visually non-obvious cues for the reasoning process. Specifically, we make utilization of caption information comprehensively from both explicit and implicit perspectives. Moreover, we further develop a mutual correlation (MC) module to learn the complex interaction of knowledge-enhanced representations, Therefore, KBCEN surpasses baseline methods on all of the three benchmark test sets, even taking into account these enhanced baselines incorporating the caption as input.

By observing top-1 accuracy and top-3 accuracy in Table 2, KBCEN significantly outperforms previous models, particularly in terms of the top-3 accuracy. In addition, it is not difficult to find some differences in the model performance on three different datasets. Compared with OK-VQA v1.0, performance on OK-VQA v1.1 has a slight improvement. As shown in Table 1, the word stemming method is adopted on the raw answers to improve their quality for OK-VQA v1.1 dataset, resulting in a more coherent answer vocabulary. While for A-OKVQA, as described in Sect. 4.1, questions in A-OKVQA are more challenging, conceptually diverse and require knowledge outside the image. They cannot be answered by simply querying the knowledge base. These may result in a decrease concerning to model accuracy.

### 4.5 Ablation performance

To evaluate which component of KBCEN is really important, we further perform an ablation study. The results are reported in Table 3.

For information utilization, we employ caption information to help understand image semantics, and introduce KG to integrate external knowledge. As shown in Table 3 (1)–(2),




(a) OK-VQA v1.0 (b) OK-VQA v1.1 (c) A-OKVQA

**Fig. 4** Sensitivity study of caption quality on KBCEN. Here, we use the division interval of the test set as the representative shown in the abscissa axis, where N represents the number of each test set. Top-1 accuracy and top-3 accuracy are calculated on each interval

the performance of KBCEN significantly decreases when they are removed separately, which indicates both semantic caption and graph knowledge are critical for the KB-VQA task.

Recalling the model architecture, as shown in Table 3 (3), we can observe that model performance significantly declines when removing the MMBERT module. Moreover, implicit and explicit knowledge-enhanced representations are mutually correlated after representing them separately. We are curious whether MC is vital for KBCEN. Thus, we remove MC to verify it. According to the results in Table 3 (4)–(5), when removing matrix multiplication and attention mechanism separately from MC, varying degrees of model performance reduction could be observed. The results in Table 3 (6) illustrate MC could build up correlations between explicit and implicit knowledge-enhanced representations.

### 4.6 Effect of caption quality

In this subsection, we conduct a sensitivity analysis to figure out how variations in caption quality might affect model performance. The results are shown in Fig. 4.

To evaluate the potential impact of caption quality on model performance, we rank the test data of three datasets based on the confidence scores of captions generated by VinVL, respectively. Subsequently, we divide the test sets into eight partitions in ascending order of confidence scores. Each partition for OK-VQA v1.0 and v1.1 contains 630 data points, and each partition for A-OKVQA consists of 837 data points. The top-1 accuracy and top-3 accuracy of the model are calculated in each partition on three datasets separately.

From the experimental results, we observe that as the caption confidence score increases, the performance of KBCEN increases slightly on all three datasets. As the quality of captions improves, the overall top-1 accuracy and top-3 accuracy of the model increase, indicating that higher-quality captions are indeed beneficial for model performance. When the caption scores are relatively low, corresponding to the first few divisions, we can see a marginal decline in model performance. The top-1 accuracy remains within a decline range of less than 2%, and the top-3 accuracy stays within a decline range of approximately 1%, which demonstrates the robustness of our model. By introducing explicit and implicit knowledge, as well as a mutual correlation (MC) to learn complex interactions of knowledge-enhanced representations, our model maintains good performance and demonstrates its robustness.

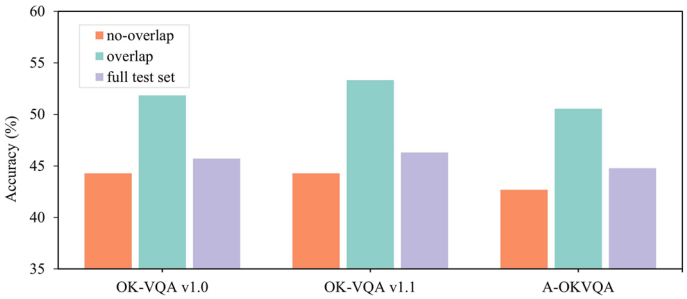



**Fig. 5** The performances of KBCEN versus the overlapping of caption and answer on the test sets of OK-VQA v1.0, v1.1 and A-OKVQA datasets, and the top-1 accuracy is calculated here

### 4.7 Statistics of overlap between answers and captions

To evaluate the overlap between answers and captions, we further conduct experiments on the OK-VQA and A-OKVQA datasets by assessing the frequency of an answer appearing within its corresponding caption. For better visualization, the changes in performance versus the overlapping of caption and answer are illustrated in Fig. 5.

On the training set and the test set of the OK-VQA v1.0 dataset, the experimental results show that the overlap frequency between answers and captions is 20.14% and 21.39%, respectively. Similarly, on the OK-VQA v1.1 dataset, the experimental results of the training set and the test set are 20.97% and 22.16%, respectively. On the A-OKVQA dataset, the experimental results of the training set and the test set are 25.12% and 26.68%, respectively. We divided the test sets into two categories: instances with no overlap between captions and answers, and instances with overlap. Then, we computed the top-1 accuracy for these divisions. From the results, we can see that captions are indeed helpful for predicting the correct answers for the KB-VQA task, especially for those where the answers appear in captions. Besides, our model is also robust for cases where the answer does not appear in captions. By better knowledge reasoning from both explicit and implicit perspectives, and leveraging the MC module to discern intricate correlations between explicit and implicit representations, our model can also reason effectively from images and questions.

### 4.8 Interpretability and case study

In the above, we have proved the effectiveness of our proposed KBCEN and its components in a quantitative manner. In addition to the effectiveness, we also conduct some experiments on the intrinsical interpretability of the caption enhanced KB-VQA system and provide case studies for the sake of intuition.

#### 4.8.1 Interpretability of KG

To illustrate the interpretability of KG, we list some examples in Fig. 6. In the first example, there is a bunch of fruits and vegetables in the image, and the question is *'What is the nutrition value of the fruits?'*. The prediction is supported by a knowledge graph edge that indicates *'fruit has_a vitamin'*, which makes the answer more likely to be *'vitamin'*. In the second example, there are many white cakes on the table and our model correctly predicts that fondant is used to make the flowers. Some knowledge triples such as *'cake related_to fondant'*


**Question:** What is the **nutrition** value of the **fruits**?
**Caption:** A bunch of **fruits** and **vegetables** on a table.
**Answer:** vitamin

| Knowledge | |
|---|---|
| (**fruit**, HasProperty, healthy) | (**fruit**, HasProperty, colorful) |
| (**vegetable**, IsA, food) | (**vegetable**, RelatedTo, **fruit**) |
| (**vitamin**, UsedFor, healthy) | (**fruit**, HasA, **vitamin**) |

**Question:** What is the **material** used to make the flowers?
**Caption:** A couple of white **cakes** sitting on top of a table.
**Answer:** fondant

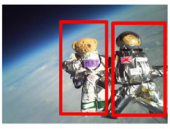
| Knowledge | |
|---|---|
| (**cake**, IsA, baked food) | (**cake**, IsA, dessert) |
| (**cake**, UsedFor, eating) | (**fondant**, HasContext, cuisine) |
| (**fondant**, HasContext, food) | (**cake**, RelatedTo, **fondant**) |

**Question:** How is this form of **transportation** powered?
**Caption:** A **green bus** is parked at a bus stop.
**Answer:** electricity

| Knowledge | |
|---|---|
| (**bus**, UsedFor, **transportation**) | (power, RelatedTo, transfer) |
| (**electrical power**, IsA, power) | (**green**, IsA, environmentalist) |
| (**bus**, IsA, **public transport**) | (**electricity**, RelatedTo, power) |

**Question:** Why does this person have **protective clothing** on?
**Caption:** A man wearing a **helmet** sitting on a red motorcycle.
**Answer:** safety

| Knowledge | |
|---|---|
| (**helmet**, IsA, a covering) | (**helmet**,RelatedTo, motorcycle) |
| (**helmet**, IsA, **protective garment**) | |
| (garment, IsA, **clothing** ) | (**protective**, RelatedTo, **safety**) |



**Question:** Where is this picture taken from?
**Caption:** Two teddy bears dressed in **space suits** on a **rocket**.
**Answer:** space

| Knowledge | |
|---|---|
| (**space suit**, RelatedTo, **space**) | (**space suit**, UsedFor, **protect**) |
| (**skyspace**,DerivedFrom, **space**) | (space rocket, IsA, **rocket**) |
| (Earth, AtLocation, **space**) | (**rocket**, RelatedTo, **space**) |

**Question:** What kind of **fuel** does this toy use?
**Caption:** A man sitting behind a small model **airplane**.
**Answer:** gasoline

| Knowledge | |
|---|---|
| (**airplane**, IsA, **fuel powered device**) | |
| (**gasoline**, UsedFor, **fueling an engine**) | |
| (**gasoline**, CapableOf, power) | (**gasoline**, IsA, **fuel**) |

**Fig. 6** Visualization of the predicted answers (left) and supporting knowledge (right) of KBCEN. We emphasize the most relevant knowledge related to the image, question, caption and answer

and *'fondant has_context food'* can thus be helpful for the answering. In the third question, it asks how the green bus is powered and the model guesses '*electricity*.' This is supported by some knowledge that indicates *'bus is_a public transport'*, *'green is_a environmentalist'* and *'electricity related_to power'*. In the next question, it wonders why the man has protective clothing on. The response is supported by certain information such as *'helmet used_for protection'* supports the answer, and there is an edge here that connects the word '*protective*' in the question directly to the answer '*safety*.' In the next example, we see that there are two teddy bears dressed in space suits on a rocket. Some knowledge triples, such as *'space suit related_to space'* and *'rocket related_to space'*, help to deduce the correct answer. In the last example, the provided knowledge establishes some links among the answer *'gasoline'*, the words *'fuel'* in the question and *'airplane'* in the caption. Furthermore, it also provides additional supplementary information like *'gasoline used_for fueling an engine'*, which corroborates the right answer.


**Fig. 7** Case study of MSG-KRM [55] and our proposed KBCEN on OK-VQA v1.1 dataset

### 4.8.2 Influence of caption

To better show the impact of the caption information, we select some examples from test sets and show predictions of our model versus the baseline MSG-KRM [55]. The results are presented in Fig. 7. In the first example, there are two giraffes sitting next to each other. MSG-KRM mispredicts the result as *'giraffe'*. We speculate the reason is that *'giraffe'* is the most salient object in the image. However, when taking caption information into consideration, it describes, from a semantic perspective, two giraffes sitting down for rest, so *'rest'* would be correctly predicted. For the second instance, two common modes of transportation are visible on the street: a truck and a bus. However, with the assistance of the caption, KBCEN can find their relationship in this situation and know that a tow truck is towing a bus down a street. In the third illustration, it is also clear from the caption that the vehicle is an ice cream truck that sells ice cream. The last illustration shows a man using a surfboard to ride a wave. *'Surfboard'* and *'ocean'* in the caption can link to the knowledge graph to find possible surfing locations, thereby narrowing down the range of the candidate answers. Combining the question words *'island'* and *'US'*, it is easy for our model to come to the correct answer '*Hawaii*.'

## 4.9 Discussion

In this section, we first discuss the future direction of integrating large language models for the KB-VQA task, and then explore potential applications of our proposed model.

### 4.9.1 Benefits from LLMs

The future direction of integrating large language models for KB-VQA tasks holds immense promise and potential for enhancing the capabilities of these systems. These large language models, with their deep understanding of natural language and vast knowledge base, can significantly improve reasoning and inference capabilities. By integrating these models, we can expect advancements in contextual understanding, commonsense reasoning and multimodal comprehension, leading to more accurate and nuanced answers to visual questions. Moreover, ongoing research into fine-tuning and customizing these models for specific domains or tasks will likely result in even greater performance gains. Thus, the future direction involves harnessing the great potential of these advanced language models to create robust, versatile and intelligent KB-VQA systems that can excel across diverse domains and applications.

### 4.9.2 Potential applications of KBCEN

Further exploration of potential applications of our proposed model could significantly enhance its utility in various domains such as education, healthcare, robotics and so on. In the field of education, the proposed model could be deployed as an interactive learning tool, allowing students to ask questions about visual content and receive accurate answers based on knowledge graphs and contextual understanding. In healthcare, the model could assist medical professionals in interpreting medical images and answering queries related to patient diagnoses and treatment plans. Moreover, in the field of autonomous systems and robotics, integrating the proposed model into robotics could enable robots to understand and respond to visual commands or questions. This could be valuable in various fields, from household robots assisting with tasks to industrial robots optimizing workflows. Overall, continual refinement and validation through interdisciplinary collaborations and real-world applications will contribute to the ongoing evolution and impact of the proposed model in diverse fields.

## 5 Conclusion

In this paper, we focused on knowledge-based visual question answering and argued that the image caption is a critical factor for KB-VQA, which can provide visually inconspicuous or invisible cues for the reasoning process. To this end, we proposed a Knowledge Based Caption Enhanced Net (KBCEN) to incorporate caption information into the KB-VQA process. Specifically, we made utilization of caption information comprehensively from both explicit and implicit perspectives for better knowledge reasoning. Moreover, we further designed a mutual correlation (MC) module to learn the correlation between explicit and implicit representations, and enhance each other in a mutual manner. Extensive experiment results on three popular and public datasets showed the superiority of our proposed method. The rationality of KBCEN could also be demonstrated by both quantitative and qualitative evaluation and analyses.

In the future, we will further study more comprehensive utilization of caption effects and conduct deeper research on leveraging external knowledge sources for better knowledge reasoning. In addition, how to effectively integrate large language models with billions of parameters and combine them with different types of knowledge bases will be an exploration direction. Last but not least, our proposed method could provide insights and inspiration for many other practical knowledge-based systems, where richer joint knowledge representation and mutual correlation for multimodal knowledge reasoning could be well applied.

**Acknowledgements** This research was partially supported by grants from the National Key Research and Development Program of China (Grant No. 2021YFF0901000), the National Natural Science Foundation of China (No. 62337001, No. 62376086), Joint Funds of the National Natural Science Foundation of China (Grant No. U22A2094) and the Fundamental Research Funds for the Central Universities.

**Author contributions** Bin Feng was involved in conceptualization, investigation, methodology, formal analysis, writing—original draft. Shulan Ruan helped in software, validation, writing—review and editing. Likang Wu contributed to data curation, investigation. Huijie Liu assisted in visualization, formal analysis. Kai Zhang was involved in investigation, supervision. Kun Zhang helped in writing—review and editing. Qi Liu assisted in writing—review and editing, resources, supervision. Enhong Chen performed project administration, supervision.

**Data availability** No datasets were generated or analyzed during the current study.

## Declarations

**Conflict of interest** The authors declare no conflict of interest.

## References


1. Marino K, Rastegari M, Farhadi A, Mottaghi R (2019) OK-VQA: a visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3195–3204
2. Kim J-H, Jun J, Zhang B-T (2018) Bilinear attention networks. In: Advances in neural information processing systems, vol 31
3. Li G, Wang X, Zhu W (2020) Boosting visual question answering with context-aware knowledge aggregation. In: Proceedings of the 28th ACM international conference on multimedia, pp 1227–1235
4. Ben-Younes H, Cadene R, Cord M, Thome N (2017) Mutan: Multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 2612–2620
5. Guo W, Zhang Y, Yang J, Yuan X (2021) Re-attention for visual question answering. IEEE Trans Image Process 30:6730–6743
6. Wang P, Wu Q, Shen C, Dick A (2018) Hengel: FVQA: Fact-based visual question answering. IEEE Trans Pattern Anal Mach Intell 40(10):2413–2427
7. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: International conference on learning representations
8. Wu J, Lu J, Sabharwal A, Mottaghi R (2022) Multi-modal answer validation for knowledge-based vqa. In: Proceedings of the AAAI conference on artificial intelligence, vol 36. pp 2712–2721
9. Marino K, Chen X, Parikh D, Gupta A, Rohrbach M (2021) Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based VQA. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14111–14121
10. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-CNN. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969
11. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, vol 28
12. Lu J, Batra D, Parikh D, Lee S (2019) Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in neural information processing systems, vol 32
13. Tan H, Bansal M (2019) Lxmert: Learning cross-modality encoder representations from transformers. 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), pp 5100–5111
14. Gardères F, Ziaeefard M, Abeloos B, Lecue F (2020) Conceptbert: Concept-aware representation for visual question answering. In: Findings of the association for computational linguistics: EMNLP 2020, pp 489–498
15. Zhu Z, Yu J, Wang Y, Sun Y, Hu Y, Wu Q (2021) Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In: Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence, pp 1097–1103
16. Kenton JDM-WC, Toutanova LK (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp 4171–4186
17. Li LH, Yatskar M, Yin D, Hsieh C-J, Chang K-W (2019) Visualbert: a simple and performant baseline for vision and language. arXiv:1908.03557
18. Malinowski M, Rohrbach M, Fritz M (2015) Ask your neurons: a neural-based approach to answering questions about images. In: Proceedings of the IEEE international conference on computer vision, pp 1–9
19. Ren M, Kiros R, Zemel R (2015) Exploring models and data for image question answering. In: Advances in neural information processing systems vol 28
20. Sharma H, Jalal AS (2022) Convolutional neural networks-based VQA model. In: Proceedings of international conference on frontiers in computing and systems: COMSYS 2021, Springer, pp 109–116
21. Wang F, Liu Q, Chen E, Huang Z, Yin Y, Wang S, Su Y (2022) NeuralCD: a general framework for cognitive diagnosis. IEEE Trans Knowl Data Eng 35(8):8312–8327
22. Yu Z, Yu J, Fan J, Tao D (2017) Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 1821–1830

23. Ben-Younes H, Cadene R, Thome N, Cord M (2019) Block: bilinear superdiagonal fusion for visual question answering and visual relationship detection. In: Proceedings of the AAAI conference on artificial intelligence, vol 33. pp 8102–8109
24. Yang Z, He X, Gao J, Deng L, Smola A (2016) Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 21–29
25. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) 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, pp 6077–6086
26. Liang J, Jiang L, Cao L, Li L-J, Hauptmann AG (2018) Focal visual-text attention for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6135–6143
27. Changpinyo S, Kukliansy D, Szpektor I, Chen X, Ding N, Soricut R (2022) All you may need for VQA are image captions. In: Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: human language technologies, pp 1947–1963
28. Wang P, Wu Q, Shen C, Dick A, Hengel A (2017) Explicit knowledge-based reasoning for visual question answering. In: Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI-17, pp 1290–1296
29. Schwenk D, Khandelwal A, Clark C, Marino K, Mottaghi R (2022) A-okvqa: A benchmark for visual question answering using world knowledge. In: Computer Vision–ECCV 2022: 17th European conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII. Springer, pp 146–162
30. Shah S, Mishra A, Yadati N, Talukdar PP (2019) Kvqa: Knowledge-aware visual question answering. In: Proceedings of the AAAI conference on artificial intelligence, vol 33. pp 8876–8884
31. Gao F, Ping Q, Thattai G, Reganti A, Wu YN, Natarajan P (2022) Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5067–5077
32. Formica A, Mele I, Taglino F (2024) A template-based approach for question answering over knowledge bases. Knowl Inf Syst 66(1):453–479
33. Lin W, Byrne B (2022) Retrieval augmented visual question answering with outside knowledge. In: Proceedings of the 2022 conference on empirical methods in natural language processing, pp 11238–11254
34. Shao Z, Yu Z, Wang M, Yu J (2023) Prompting large language models with answer heuristics for knowledge-based visual question answering. In: Computer vision and pattern recognition (CVPR)
35. Lin Y, Xie Y, Chen D, Xu Y, Zhu C, Yuan L (2022) Revive: Regional visual representation matters in knowledge-based visual question answering. In: Advances in neural information processing systems
36. Rathnayake H, Sumanapala J, Rukshani R, Ranathunga S (2022) Adapter-based fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification. Knowl Inf Syst 64(7):1937–1966
37. Yang Z, Gan Z, Wang J, Hu X, Lu Y, Liu Z, Wang L (2022) An empirical study of GPT-3 for few-shot knowledge-based vqa. In: Proceedings of the AAAI conference on artificial intelligence, vol 36. pp 3081–3089
38. Huang D, Wei Z, Yue A, Zhao X, Chen Z, Li R, Jiang K, Chang B, Zhang Q, Zhang S et al (2023) Dsqa-llm: Domain-specific intelligent question answering based on large language model. In: International conference on AI-generated content, Springer, pp 170–180
39. Yu Z, Ouyang X, Shao Z, Wang M, Yu J (2023) Prophet: Prompting large language models with complementary answer heuristics for knowledge-based visual question answering. arXiv:2303.01903
40. Hu Y, Hua H, Yang Z, Shi W, Smith NA, Luo J (2022) Promptcap: prompt-guided task-aware image captioning. arXiv:2211.09699
41. Gui L, Wang B, Huang Q, Hauptmann A, Bisk Y, Gao J (2021) Kat: a knowledge augmented transformer for vision-and-language. arXiv:2112.08614
42. Li S, Luo C, Zhu Y, Wu W (2023) Bold driver and static restart fused adaptive momentum for visual question answering. Knowl Inf Syst 65(2):921–943
43. Muscetti M, Rinaldi AM, Russo C, Tommasino C (2022) Multimedia ontology population through semantic analysis and hierarchical deep features extraction techniques. Knowl Inf Syst 64(5):1283–1303
44. Gao J, Al-Sabri R, Oloulade BM, Chen J, Lyu T, Wu Z (2023) Gm2nas: multitask multiview graph neural architecture search. Knowl Inf Syst 65(10):4021–4054
45. Su Z, Gou G (2024) Knowledge enhancement and scene understanding for knowledge-based visual question answering. Knowl Inf Syst 66(3):2193–2208
46. Ruan S, Zhang Y, Zhang K, Fan Y, Tang F, Liu Q, Chen E (2021) Dae-gan: dynamic aspect-aware gan for text-to-image synthesis. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 13960–13969

47. Zhou B, Tian Y, Sukhbaatar S, Szlam A, Fergus R (2015) Simple baseline for visual question answering. arXiv:1512.02167
48. Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M (2016) Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 457–468
49. Gao P, Jiang Z, You H, Lu P, Hoi SC, Wang X, Li H (2019) Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6639–6648
50. Hannan D, Jain A, Bansal M (2020) Manymodalqa: Modality disambiguation and qa over diverse inputs. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 7879–7886
51. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inform Process Syst 30
52. Singh LG, Singh SR (2024) Sentiment analysis of tweets using text and graph multi-views learning. Knowl Inform Syst. https://doi.org/10.1007/s10115-023-02053-8
53. Ding Y, Yu J, Liu B, Hu Y, Cui M, Wu Q (2022) Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5089–5098
54. Salaberria A, Azkune G, Lacalle OL, Soroa A, Agirre E (2023) Image captioning for effective use of language models in knowledge-based visual question answering. Expert Syst Appl 212:118669
55. Jiang L, Meng Z (2023) Knowledge-based visual question answering using multi-modal semantic graph. Electronics 12(6):1390
56. Schelling B, Plant C (2020) Dataset-transformation: improving clustering by enhancing the structure with dipscaling and diptransformation. Knowl Inf Syst 62(2):457–484
57. Wang M, Zhou X, Chen Y (2024) JMFEEL-NET: a joint multi-scale feature enhancement and lightweight transformer network for crowd counting. Knowl Inform Syst. https://doi.org/10.1007/s10115-023-02056-5
58. Li G, Duan N, Fang Y, Gong M, Jiang D (2020) Unicoder-vl: a universal encoder for vision and language by cross-modal pre-training. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 11336–11344
59. Su W, Zhu X, Cao Y, Li B, Lu L, Wei F, Dai J (2019) Vl-bert: Pre-training of generic visual-linguistic representations. In: International conference on learning representations
60. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, PMLR, pp 8748–8763
61. Li J, Li D, Xiong C, Hoi S (2022) Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning, PMLR, pp 12888–12900
62. Girdhar R, El-Nouby A, Liu Z, Singh M, Alwala KV, Joulin A, Misra I (2023) Imagebind: one embedding space to bind them all. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 15180–15190
63. Zhang P, Li X, Hu X, Yang J, Zhang L, Wang L, Choi Y, Gao J (2021) Vinvl: revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5579–5588
64. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L-J, Shamma DA et al (2017) Visual genome: connecting language and vision using crowdsourced dense image annotations. Int J Comput Vision 123(1):32–73
65. Vinyals O, Toshev A, Bengio S, Erhan D (2016) Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE Trans Pattern Anal Mach Intell 39(4):652–663
66. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, PMLR, pp 2048–2057
67. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K et al (2016) Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144
68. Liu H, Singh P (2004) Conceptnet-a practical commonsense reasoning tool-kit. BT Technol J 22(4):211–226
69. Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z (2007) Dbpedia: a nucleus for a web of open data. In: International semantic web conference, Springer, pp 722–735
70. Bhakthavatsalam S, Richardson K, Tandon N, Clark P (2020) Do dogs have whiskers? a new knowledge base of haspart relations. arXiv:2006.07510

71. Schlichtkrull M, Kipf TN, Bloem P, Berg Rvd, Titov I, Welling M (2018) Modeling relational data with graph convolutional networks. In: European semantic web conference. Springer, pp 593–607
72. Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. In: International conference on learning representations
73. Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D (2017) Making the v in vqa matter: elevating the role of image understanding in visual question answering. In: Proceedings of the ieee conference on computer vision and pattern recognition, pp 6904–6913
74. Ruan S, Zhang K, Wu L, Xu T, Liu Q, Chen E (2021) Color enhanced cross correlation net for image sentiment analysis. IEEE Trans Multim. https://doi.org/10.1109/TMM.2021.3118208
75. Sun R, Tao H, Chen Y, Liu Q (2024) HACAN: a hierarchical answer-aware and context-aware network for question generation. Front Comput Sci 18(5):185321
76. Guo D, Xu C, Tao D (2023) Bilinear graph networks for visual question answering. IEEE Trans Neural Netw Learn Syst 34(2):1023–1034. https://doi.org/10.1109/TNNLS.2021.3104937
77. Mishra A, Anand A, Guha P (2023) Dual attention and question categorization-based visual question answering. IEEE Trans Artif Intell 4(1):81–91. https://doi.org/10.1109/TAI.2022.3160418
78. Song L, Li J, Liu J, Yang Y, Shang X, Sun M (2023) Answering knowledge-based visual questions via the exploration of question purpose. Pattern Recogn 133:109015



**Bin Feng** received the B.E. degree from University of Science and Technology of China, Hefei, China, in 2018. He is currently working toward the Ph.D. degree with the School of Computer Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include knowledge discovery, visual question answering and multimodal applications.

**Shulan Ruan** received the Ph.D degree in computer science and technology from University of Science and Technology of China, Hefei, China, in 2024. He is currently a postdoctoral fellow in Shenzhen International Graduate School at Tsinghua University. His research interests include computer vision, data mining and multimodal modeling. He has published over 10 papers in ICCV, AAAI, IJCA, TMM, etc.

**Likang Wu** is currently a Ph.D. candidate in the School of Computer Science and Technology at University of Science and Technology of China, Hefei, China. His major research interests include data mining, graph embedding and knowledge graph. He has published several papers in refereed journals and conference proceedings, such as Pattern Recognition, IEEE TKDD, ACM SIGKDD, AAAI, IJCAI, ACM SIGIR and DASFAA.

**Huijie Liu** received the B.S. degree in statistics from the University of Science and Technology of China, Hefei, China, in 2019. She is currently pursuing the Ph.D. degree in the Anhui Province Key Laboratory of Big Data Analysis and Application, School of Data Science, University of Science and Technology of China, Hefei, China. Her current research interests include network embedding and its application for knowledge discovery and data mining.

**Kai Zhang** is an associate researcher at the University of Science and Technology of China (USTC). He received his Ph.D. degree in Computer Science from USTC. His general area of research is natural language processing and knowledge discovery. He has published prolifically in refereed journals and conference proceedings, e.g., TKDE, WWWJ, ACL, IJCAI, AAAI, KDD, SIGIR and WWW. He has served regularly on the program committees of a number of conferences and is a reviewer for the leading academic journals in his fields. He is a member of ACM, SIGIR, AAAI and CCF. He was awarded the CCML 2019 Best Student Paper Award as the first author.

**Kun Zhang** received the Ph.D. degree in computer science and technology from University of Science and Technology of China, Hefei, China, in 2019. He is currently a faculty member with the Hefei University of Technology (HFUT), China. His research interests include natural language processing, recommendation system and text mining. He has published several papers in refereed conference proceedings such as AAAI, KDD and ICDM. He received the KDD 2018 Best Student Paper Award.

**Qi Liu** received the Ph.D. degree from University of Science and Technology of China (USTC), Hefei, China, in 2013. He is currently a Professor in the School of Computer Science and Technology at USTC. His general area of research is data mining and knowledge discovery. He has published prolifically in referred journals and conference proceedings (e.g., TKDE, TOIS, TKDD, TIST, KDD, IJCAI, AAAI, ICDM, SDM and CIKM). He is an Associate Editor of IEEE TBD and Neurocomputing. He has served regularly in the program committees of a number of conferences, and is a reviewer for the leading academic journals in his fields. He is a member of ACM and IEEE. He was the recipient of KDD'18 Best Student Paper Award and ICDM'11 Best Research Paper Award. He was also the recipient of China Outstanding Youth Science Foundation in 2019.

**Enhong Chen** received the Ph.D. degree from University of Science and Technology of China (USTC). He is a professor, executive dean of the School of Data Science and vice dean of the School of Computer Science, USTC. He is a CCF Fellow and ACM distinguished member. His general area of research includes data mining and machine learning, social network analysis and recommender systems. He has published more than 100 papers in refereed conferences and journals, including the IEEE Transactions on Knowledge and Data Engineering, the IEEE Transactions on Mobile Computing, KDD, ICDM, NIPS and CIKM. He was on program committees of numerous conferences including KDD, ICDM and SDM. His research is supported by the National Science Foundation for Distinguished Young Scholars of China. He is an IEEE Fellow.