Dynamic Key-Value Memory Enhanced Multi-Step Graph Reasoning for Knowledge-Based Visual Question Answering

Mingxiao Li[1], Marie-Francine Moens[1]

[1] KU Leuven
mingxiao.li@kuleuven.be, sien.moens@cs.kuleuven.be

Abstract

Knowledge-based visual question answering (VQA) is a vision-language task that requires an agent to correctly answer image-related questions using knowledge that is not presented in the given image. It is not only a more challenging task than regular VQA but also a vital step towards building a general VQA system. Most existing knowledge-based VQA systems process knowledge and image information similarly and ignore the fact that the knowledge base (KB) contains complete information about a triplet, while the extracted image information might be incomplete as the relations between two objects are missing or wrongly detected. In this paper, we propose a novel model named dynamic knowledge memory enhanced multi-step graph reasoning (DMMGR), which performs explicit and implicit reasoning over a key-value knowledge memory module and a spatial-aware image graph, respectively. Specifically, the memory module learns a dynamic knowledge representation and generates a knowledge-aware question representation at each reasoning step. Then, this representation is used to guide a graph attention operator over the spatial-aware image graph. Our model achieves new state-of-the-art accuracy on the KRVQR and FVQA datasets. We also conduct ablation experiments to prove the effectiveness of each component of the proposed model.

Introduction

Over the past few years, the domain of visual question answering (VQA) (Antol et al. 2015) has attracted great attention and witnessed significant progress (Antol et al. 2015; Lu et al. 2016; Hudson and Manning 2019). However, most VQA models cannot answer questions that require external knowledge beyond what is provided in the image. Considering the top example in Figure 1, the question is "What is the relation between the object that belongs to the category of eukaryotes and the fork in the image?". To correctly answer this question, it is necessary to both understand the visible content in the image and incorporate the external knowledge that a cucumber belongs to the biological category of eukaryotes. When facing such challenging questions, we humans can easily combine the image content with general knowledge that is required for answering this question, while many current VQA models fail due to their incapability to utilize external knowledge.

Figure 1: Two examples taken from the KRVQR dataset. To correctly answer the question in the image, the model should be able to utilize supporting facts which cannot be seen in the image.

To bridge this gap between human behavior and current VQA models, several knowledge-based VQA datasets have been proposed. Wang et al. (2017) introduced the "fact-based VQA (FVQA)" task and developed the first KVQA dataset containing images, questions with answers and a knowledge base (KB) of fact triplets extracted from different sources including ConceptNET (Speer, Chin, and Havasi 2017), WebChild (Tandon et al. 2014) and DBPedia (Auer et al. 2007). Recently, Cao et al. (2021) introduced the first large-scale knowledge-based VQA dataset: Knowledge-routed VQA (KRVQR) that contains a KB for answering the questions. In this work, we mainly focus on the KRVQR dataset and also test our model on the FVQA dataset. Other VQA datasets that require external knowledge exist (Marino et al. 2019; Jain et al. 2021) but here the task is to search for external knowledge, which is not the scope of this work.

Graph-based approaches (Narasimhan, Lazebnik, and Schwing 2018; Zhu et al. 2020) have achieved successes in the field of knowledge VQA. They use one or several graphs to represent the information sources and conduct cross-graph learning or different fusion methods to infer the answer to the given question. However, they either suffer from the drawbacks of ignoring the dynamics of multi-step

reasoning, or from performing an identical reasoning procedure for both the knowledge facts and the image. Such approaches neglect that the knowledge facts provide the explicit information of a triplet, while the image graph contains only implicit information about the image.

In this work, we propose the DMMGR or **D**ynamic knowledge **M**emory enhanced **M**ulti-step **G**raph **R**easoning model, which performs explicit and implicit reasoning over a KB and a spatial-aware image graph, respectively. Specifically, we see the reasoning over the knowledge base as a problem of performing key addressing and value reading over a key-value memory and propose a novel dynamic key-value knowledge memory module to learn a question-aware knowledge representation at each reasoning step. This is different from previous key-value memory networks (Miller et al. 2016; Xu et al. 2019) whose key is the subject and relation of a triplet and value is the object.Our proposed module dynamically learns a question representation that can reason about the subject, relation and object of a knowledge triplet. We depict the image as a spatial-aware image graph where the nodes are the embeddings of the objects detected using Faster-RCNN (Ren et al. 2015) and the edges are embeddings of their relative positions. Inspired by (Gu et al. 2019; Zareian, Karaman, and Chang 2020) who leveraged common sense knowledge for scene graph generation, we use a knowledge-aware question representation, which is learned by applying explicit reasoning over the knowledge memory. This question representation performs implicit reasoning over the spatial-aware image graph. DMMGR thus implements multi-step reasoning by iteratively performing explicit reasoning over the dynamic knowledge memory and implicit reasoning over the spatial-aware image graph.

In summary, the main contributions of this paper are as follows, (1) We propose a novel dynamic knowledge memory module that learns a representation of knowledge triplets and generates a knowledge-aware question representation. (2) We introduce a question and knowledge guided graph reasoning module, where we use the representation of related knowledge triplets to guide the reasoning over a sparse spatial-aware image graph. (3) We perform an ablation study to verify the contribution of each model component, and attention visualization shows that our model has good interpretability.

Related Work

Visual Question Answering

The VQA task, where a VQA agent is expected to correctly answer a question related to an image, was proposed by Antol et al. (2015). Most of the early VQA models (Antol et al. 2015; Andreas et al. 2016; Ben-younes et al. 2017; Fukui et al. 2016; Lu et al. 2016; Ma, Lu, and Li 2016) integrate a CNN-RNN based architecture that fuses the RNN encoding of the question and the CNN encoding of the image to predict the answer, possibly improved by attention mechanisms to highlight the visual objects that are related to the question (Yang et al. 2016; Anderson et al. 2018; Lu et al. 2016). Recently, graph neural networks that represent the image as a scene graph, where nodes are objects and edges are relations

between two connected objects, has attracted attention in many vision-language tasks including VQA. Teney, Liu, and van den Hengel (2017) represented both a language question and an image as two graphs and applied graph attention and an aggregation operator to infer the corresponding answer. Hu et al. (2019) and Wang et al. (2019) used question guided image graph attention to generate a question-aware image graph representation. Different from the aforementioned graph models, Norcliffe-Brown, Vafeias, and Parisot (2018) proposed a graph learner, where a fully connected image graph whose nodes are region features, and edges represent position information was constructed based on the given question, and this model performs graph convolution and max pooling to predict the answer.

Knowledge-Based VQA

Knowledge-based VQA (KVQA) requires the model to use knowledge outside the image to answer questions correctly. Compared to the original VQA task, KVQA is relatively less explored (e.g. Wang et al. (2017); Marino et al. (2019)) Narasimhan, Lazebnik, and Schwing (2018) have first introduced a graph based approach to KVQA and apply a graph convolution to the fact graph to infer the answer, while Ziaeefard and Lécué (2020) use graph attention and multimodal fusion to reason over both the image graph and fact graph. Another work (Zhu et al. 2020) depicts an image as three graphs: a semantic graph built on the results of dense captioning of image regions, a fact graph representing relevant knowledge triplets, and a fully connected image graph where nodes represent region features, and iteratively perform question guided inter- and intra-graph attention to answer the question. In contrast, we use a dynamic key-value memory to represent the triplets and use a sparse spatial-aware image graph whose nodes are the object category embeddings, and edges are the relative position embeddings of two objects. As an extension of our proposed model, we have evaluated the integration of a semantic graph based on dense captioning of the image regions.

Key-Value Memory Networks

A key-value memory network (Miller et al. 2016) is seen as an extension of a memory network, (Weston, Chopra, and Bordes 2014; Sukhbaatar et al. 2015). However, different from memory networks, key-value memory networks save context as key-value pairs and split the reasoning process into key addressing and value reading. Key-value memory networks have been widely used in knowledge triplet-based question answering (Miller et al. 2016; Xu et al. 2019). For a knowledge triplet $< subject, relation, object >$, a key-value memory network saves the subject and relation as a key and the object as a value, which restricts its usage. In this paper, we introduce a dynamic key-value memory module whose key is composed of all the information of a triplet and whose value is a question-aware triplet representation. Such a memory structure is flexible enough so that it can reason about not only the object but also the subject and relation of a knowledge triplet.
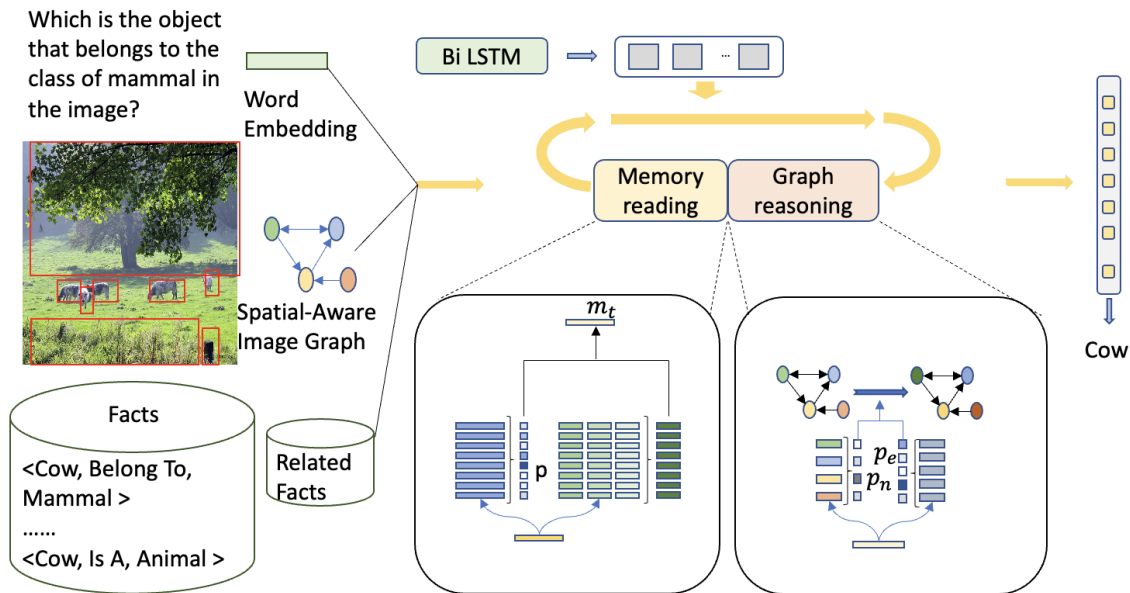
Figure 2: The architecture of our proposed DMMGR model. The module specialized in iterative reasoning consists of two components: Dynamic memory reading aimed at reading question related knowledge from the retrieved knowledge facts and resulting in a knowledge-aware question representation, and the spatial-aware image graph reasoning which performs a question and knowledge guided graph attention operation to infer the image information that is required to answer the question.

Methodology

Given a question Q, an image I and a KB K that consists of a set of facts $f_1, f_2, ..., f_n$, a KVQA model aims to predict the answer A to Q by reasoning over the image and the KB. Each fact is represented as a resource description framework (RDF) triplet of the form $f_i = (e_1, r, e_2)$, where e_1, e_2 are entities and r is the relation between e_1 and e_2. In our model, we first conduct explicit reasoning over the dynamic key-value memory module to extract question related knowledge information. Then we use both extracted knowledge information and the question representation to perform implicit reasoning over the spatial-aware image graph. The complete model is constructed by stacking both the explicit key-value memory reading module and the implicit spatial-aware image graph reasoning module. Figure 2 shows the details of our proposed model.

Key-Value Memory Construction

Fact Retrieval. To retrieve the relevant facts f_{rel} from the KB, we detect the nouns mentioned in the question and the objects in the image using the Stanza natural language processing (NLP) tool (Qi et al. 2020) and the pretrained Faster-RCNN with RestNet-101 backbone (Ren et al. 2015), respectively. We then use a pretrained GloVe word embedding (Pennington, Socher, and Manning 2014) to represent each entity and relation in the knowledge triplet, the nouns in the question and the detected image objects, and sort the triplets based on the average cosine similarity between every word in a triplet and the nouns in the question and detected image objects ignoring pairs with zero average similarity. Finally,

the top $k = 5$ facts with the highest average cosine similarity values are retained for predicting the answer to the question.
Memory Construction. A factual triplet consists of $<$ subject, relation, object $>$, such as $<$ mouse, related to, keyboard $>$. We store the extracted triplets in a key-value memory structure. This is different from previous works (Miller et al. 2016; Xu et al. 2019) that take the subject and relation as a key and the object as a value, which strongly limits their usage in reasoning. We use the average embedding of subject, relation and object as key, and the value contains each element of a triplet: $\{F_f : [F_s, F_r, F_o]\}$, where F_f is the mean GloVe embedding of the words that form an element of the triplet, and F_s, F_r, F_o are the embeddings of the subject, relation and object, respectively. Our proposed key-value structure is capable of providing all information including subject, relation and object of a triplet.

Visual Spatial-aware Graph Construction

Given an image I, we detect objects in the image and keep the top $r = 36$ detected objects $O = \{o_i\}_{i=1}^r$. Each object is associated with a label representation $v_i \in R^{d_v}(d_v = 300)$ which is the average GloVe embedding (Pennington, Socher, and Manning 2014) of the predicted object category, and a spatial vector $b_i \in R^{d_b}(d_b = 4)$ consisting of the coordinates of the top-left and bottom-right corners of the bounding box. The previous work (Zhu et al. 2020; Hu et al. 2019) use a fully connected spatial-aware image graph, which not only increases the computational cost but also introduces noisy relations between objects, we construct a sparse and $spatial - aware$ image graph $G^V = \{V^V, E^V\}$ over

the objects O by connecting only one object with its top-5 nearest neighbors. The distance between two objects is the squared distance between the centers of their two bounding boxes. Figure 3 shows an example of how we create this spatial-aware image graph. Each node in the node set $V^V = \{v_i^V\}_{i=1}^M$ corresponds to a detected object o_i represented by its label representation, and the edges $e_{ij}^V \in E^V$ denote the relations between pairs of objects. We use a 5-dimensional relative spatial vector $r_{ij}^V = [\frac{x_i^c - x_j^c}{\sqrt{w_i h_i}}, \frac{y_i^c - y_j^c}{\sqrt{w_i h_i}}, \frac{w_j}{w_i}, \frac{h_j}{h_i}, \frac{w_j h_j}{w_i h_i}]$, where x_i^c, y_i^c are the center coordinates and w_i, h_i are the width and height of the bounding box of object i, respectively, to encode edge features in the graph.
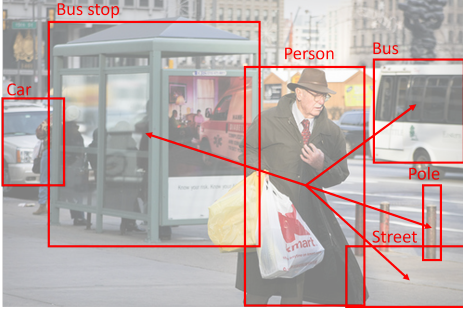

Figure 3: An example of a spatial-aware image graph. We only show the connections between person and its neighbours. There is no connection between person and car as the car is far from the person.

Iterative Reasoning Module

Algorithm 1 illustrates the flow of our proposed iterative reasoning module. The ; in algorithm 1 denotes the concatenation operation, and T is the total number of reasoning steps. The details of each module are introduced below.

Question Processing. To enhance the model's ability to perform multi-step reasoning, we generate a question vector for each reasoning step t (where $t = 1, \cdots, T$). Specifically, the given question Q of length S is converted into a sequence of GloVe word embeddings that are further processed by a 2-layer bi-directional long short-term memory (LSTM) module to generate a sequence of contextually aware word representations:

$$[h_1, h_2, \cdots, h_S] = BiLSTM(Q) \tag{1}$$

where $h_s = [\overrightarrow{h_s}; \overleftarrow{h_s}]$ is the concatenation of the forward and backward hidden state of words at position s from the last layer of the bi-directional LSTM output. At each reasoning step t, context vector c^t attends over the sequence of question words, and the question representation q^t is obtained as follows:

$$\alpha_s^t = softmax(W_1(h_s \odot (W_2^t ReLU(W_3 c^t)))) \tag{2}$$

$$q^t = \sum_{s=1}^{S} \alpha_s^t h_s \tag{3}$$

where W_1, W_2^t (as well as $W_3, \cdots, \omega_v^t, \omega_e^t$ mentioned below) are learned parameters, and $\odot$ denotes element-wise multiplication. Parameters with t superscript are learned separately for each iteration, while those without t superscript are shared across iterations. The q^t can be seen as the reasoning step-aware question representation at step t. In the first step, the reasoning step-aware vector: $c^1 = [\overrightarrow{h}_S; \overleftarrow{h}_1]$, which is the concatenation of the final hidden states obtained from the forward and backward LSTM passes. From the second step to the Tth step, c^t is evaluated as:

$$c^t = W_4^t(ReLU(W_5[R^{t-1}; I^{t-1}])) \tag{4}$$

where R^{t-1}, I^{t-1} are the knowledge-aware question representation and visual representation at step $t-1$, respectively, which are explained in details below.

Key Addressing and Value Reading. Key addressing is a process that involves finding the most relevant knowledge triplet for a given question. Following previous work (Xu et al. 2019), we compute the relevance probability p_i between the question representation q and each key representation k as:

$$\hat{q} = ReLU(W_6(ReLU(W_7 q)) \tag{5}$$

$$\hat{k}_i = ReLU(W_8(ReLU(W_9 k_i)) \tag{6}$$

$$p_i = softmax(\hat{q} \cdot \hat{k}_i^T) \tag{7}$$

As each value of our key-value memory module contains all three elements (subject, relation and object) of a triplet, we further apply an attention mechanism to compute the question-guided value embedding $\hat{t}_i$ for each value:

$$\hat{t}_{ij} = ReLU(W_{10}(ReLU(W_{11} t_{ij})) \tag{8}$$

$$s_{ij} = (1 - softmax(\hat{q} \cdot \hat{t}_{ij}^T))/2.0 \tag{9}$$

$$\hat{t}_i = \sum_{j=1}^{J} s_{ij} \hat{t}_{ij} \tag{10}$$

where $J = 3$ is the number of elements of a triplet; s_{ij} is the attention probability of each component in a knowledge triplet, and the divisor 2.0 is a normalization factor (normalize the sum of s_{ij} to one) The value of memory m^t is then computed by taking the weighted sum over the question-guided value embeddings using the relevance probabilities; this value is further used to reason over the spatial-aware image graph.

$$m^t = \sum_{i=1}^{K} p_i \hat{t}_i \tag{11}$$

where K is the number of knowledge triplets stored in the key-value memory module.

Question and Knowledge Guided Node Attention. We first merge the question information and the output of the key-value memory module to generate a knowledge-aware question representation.

$$R^t = W_{11}^t(ELU(W_{12}[q^t; m^t])) \tag{12}$$

where q^t and m^t are the question and memory embeddings generated during the $t-th$ reasoning iteration, respectively,

and ELU is the exponential linear unit activation function used here to avoid the dead neuron problem. We then use an attention mechanism to compute the relevance of each node of the spatial-aware image graph corresponding to the knowledge-aware question representation. The relevance scores are evaluated as:

$$\alpha_i = softmax(\omega_v tanh(W_{13}v_i + W_{14}R^t)) \qquad (13)$$

where v_i is the node representation of the spatial-aware image graph.

Question and Knowledge Guided Edge Attention We apply the same attention mechanism to compute the importance between edge e_{ij} and the knowledge-aware question representation R^t, which is evaluated as:

$$\beta_{ij} = softmax(\omega_e tanh(W_{15}e_{ij} + W_{16}R^t) \qquad (14)$$

Multi-head Spatial-aware Image Graph Attention. Based on the node and edge attention weights computed in Eq. 13 and Eq. 14, respectively, the node representations of the spatial-aware image graph are updated with multi-head graph attention (Veličković et al. 2017).

$$m_i^k = \sum_{j \in \mathcal{N}_i} ([\alpha_j W_{17}v_j; \beta_{ij} W_{18}e_{ij}]) \qquad (15)$$

$$h_i^k = \alpha_i ReLU(W_{19}[m_i^h; W_{20}v_i]) \qquad (16)$$

$$\hat{v}_i = LayerNorm(ELU(W_{21}[h_i^1; h_i^2; \cdots; h_i^H])) \quad (17)$$

where $\mathcal{N}_i$ is the neighborhood set of node v_i and H is the number of heads. Once all nodes are updated, the max pooling operation is conducted over all nodes to obtain the visual representation at reasoning step t:

$$I^t = MaxPooling(\{v_i\}_{i=1}^M) \qquad (18)$$

Final Prediction. We iteratively perform knowledge key addressing, value reading and spatial-aware image graph attention reasoning for T steps. At step T the given knowledge-aware question representation R^T and visual representation I^T are concatenated and processed by a 2-layer linear transformation to predict the answer to the question. During training, we simply use the cross-entropy loss function (see equation below), to optimize the differences between the predicted answer and ground truth answer.

$$L = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^L y_c log(\hat{y}_c) \qquad (19)$$

where N and L are the number of training samples and candidate answers, respectively. y_c is the ground truth answer, and $\hat{y}_c$ is the predicted answer.

Experiments

Datasets. In this paper we mainly focus on the KRVQR (Cao et al. 2021) dataset and also test our model on the FVQA (Wang et al. 2017) dataset, as these are the only two knowledge VQA datasets that provide the knowledge base for answering the questions. The KRVQR dataset consists of 32910 images and 157201 question answer pairs, divided into training, validation and test sets with proportions of

60%, 20% and 20%, respectively. The FVQA dataset contains 2190 images and 5826 questions, which are further split into training (2927) and test (2899) sets. The KRVQR dataset contains 43.5% one-step reasoning and 56.5% two-step reasoning questions, and the FVQA dataset contains only one-step reasoning questions (Wang et al. 2017). Two-step reasoning questions need to reason over two relations to infer the answer, while one-step reasoning questions need only one relation, where relations can be found in the KB and/or image. An example of a two-step and an example of a one-step reasoning question are presented on the top and bottom of Figure 1, respectively.

Evaluation Metrics. Following the literature (Cao et al. 2021; Wang et al. 2017), we evaluate our model using top-1 accuracy (KRVQR and FVQA datasets) and top-3 accuracy (FVQA dataset).

Implementation Details. We implement our model using the PyTorch framework (Paszke et al. 2019). The hidden size of the LSTM encoder is set to 512, and the dropout rate is 0.1. We set the sizes of the dynamic key-value memory embeddings and graph node and edge embeddings to 300 and 1024, respectively. The number of reasoning steps is set to 2 as the questions in the dataset require maximum 2 reasoning steps. All these parameters are selected based on the validation results. The model is trained using the Adam algorithm (Kingma and Ba 2014) with a base learning rate of $1e^{-4}$. We gradually increase the learning rate over the first two epochs and start decaying the learning rate at epoch 20. The best model is trained for approximately 40 epochs with a batch size of 128.

Results

Table 1 illustrates the accuracy comparison between our DMMGR model with other models including the state-of-the-art VQA and knowledge based VQA models. The Q-type (Cao et al. 2021), LSTM (Cao et al. 2021), FiLM (Perez et al. 2018), MFH (Yu et al. 2018), UpDown (Anderson et al. 2018), and state-of-the-art VQA model MCAN (Yu et al. 2019) were reimplemented and tested on the KRVQR dataset in the work of Cao et al. (2021), and we copy the obtained results here. As running software is not publicly available, we reimplement the Mucko model and test it on the KRVQR dataset. All our results are average over 5 runs. As presented in Table 1, our DMMGR model significantly outperforms all the other models and surpasses the current state-of-the-art model (KM-net) by approxiFmately **6%** in terms of accuracy. We also present the results of DMMGR extended with a semantic graph based on dense captioning of image regions that aims at including attributes of the image regions (DMMGR + Dense Captioning) following the method described in Yu et al. (2020). We see a slight accuracy improvement when including the dense captioning information in the DMMGR model, however, statistical significance testing (t-test $p = 0.051$) shows that no significant differences can be observed. We report the results on the FVQA dataset in Table 2. GRUC (Yu et al. 2020) is the current state-of-the-art model for the FVQA dataset. The complete GRUC model also integrates information from a pretrained image dense caption generation model. The table

Algorithm 1: Iterative Reasoning Module

Input:
    Question Q, key-value memory M, spatial-aware image graph G

Output:
    Answer prediction P

1: Process Q based on Equation 1
2: Initialize $c^1 = [\overrightarrow{h_S}; \overleftarrow{h}_1]$
3: **for** $t = 1$ to T **do**
4:     Obtain q^t based on Equations 2 and 3
5:     Perform key addressing and value reading to obtain the representation of the knowledge triplets m^t based on Equations 5$\sim$11
6:     Obtain the knowledge-aware question representation R^t based on Equation 12
7:     Compute the question knowledge-guided visual node and edge attention based on Equations 13 and 14
8:     Update the visual node representation v_i based on Equations 15$\sim$17
9:     Obtain the spatial-aware image graph output I^t based on Equation 18
10:     **if** $t < T - 1$ **then**
11:         Update context representation c^t based on Equation 4
12:     **end if**
13: **end for**
14: $P = Linear(Linear([R^T; I^T]))$
15: **return** P

shows that both the DMMGR and DMMGR + Dense Captioning models outperform GRUC, thus obtaining new state-of-the-art performance for the FVQA dataset. Integrating the dense captioning information in the DMMGR model could significantly improve its accuracy on the FVQA dataset by around 2.6% (t-test $p = 0.02$), which is different from what we observe in the KRVQR dataset. This difference might come from the difference of data distribution and types of questions in the two datasets.

Model	Accuracy
Q-type (Cao et al. 2021)	8.12
LSTM (Cao et al. 2021)	8.81
FiLM (Perez et al. 2018)	16.89
MFH (Yu et al. 2018)	19.55
UpDown (Anderson et al. 2018)	21.85
MCAN (Yu et al. 2019)	22.23
Mucko (Zhu et al. 2020)	24.00
KM-net (Cao et al. 2019)	25.19
DMMGR (2-steps)	31.4
DMMGR+Dense Captioning (2-steps)	31.8

Table 1: Top-1 accuracy comparisons among different models on the KRVQR dataset.

Model	Accuracy	
	top-1	top-3
FVQA (Ensemble) (Wang et al. 2017)	58.76	
$STTF$[1]	62.20	75.60
OB[2]	69.35	80.25
Mucko (Zhu et al. 2020)	73.06	85.94
GRUC (Yu et al. 2020)	79.63	91.20
GRUC (without Semantic graph)	78.05	87.70
DMMGR (1-step)	78.6	90.6
DMMGR+Dense Captioning (1-step)	81.20	95.38

Table 2: Top-1 and top-3 accuracy of the different models obtained on the FVQA dataset.(1:Narasimhan and Schwing (2018),2:Narasimhan, Lazebnik, and Schwing (2018))

Ablation Study

To verify the effectiveness of each component of our model, we conduct an extensive ablation study of the results obtained on the KRVQR dataset, which is the most challenging dataset (Table 3, 4 and 5). They reveal the impact of the iterative reasoning module, the dynamic key-value memory module and the knowledge-guided graph reasoning.

Does the Number of Reasoning Steps Matter?

We first verify the contribution of the iterative reasoning module by performing experiments using the DMMGR model with different numbers of reasoning steps. From Table 3, we can observe that the DMMGR model with two reasoning steps has the best performance, which is slightly higher than that of the DMMGR model with only one reasoning step. However, having more than two reasoning steps, such as three and four, dramatically decreases the accuracy by more than 4%. This is not surprising, as the KRVQR dataset only contains one-step reasoning (43.5%) and two-step (56.5%) reasoning questions.

Model (DMMGR)	Accuracy
1 step	30.5
2 steps	31.4
3 steps	27.1
4 steps	26.2

Table 3: Results in terms of top-1 accuracy of the DMMGR model obtained on the KRVQR dataset considering different numbers of reasoning steps.

Does the Proposed Key-Value Memory Really Help?

We next study the advantage of using a dynamic key-value memory module (Table 4). The experiments are conducted by replacing the memory module of the DMMGR model with different types of memory modules and testing the new models on the KRVQR dataset. The first model in Table 4 integrates a simple memory module where each slot is the average embedding of a knowledge triplet. The second
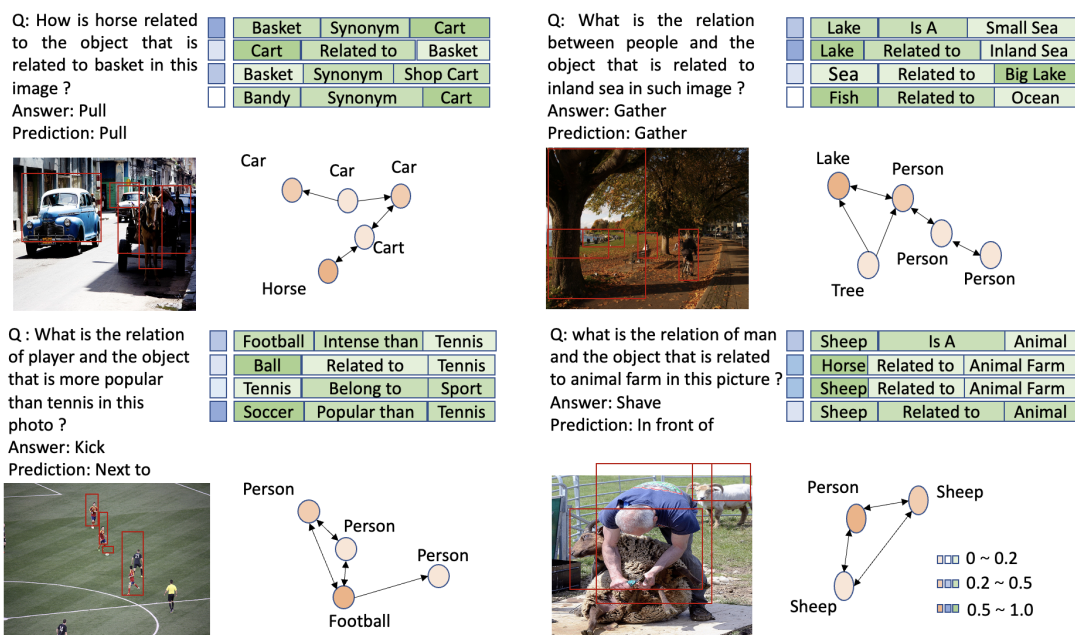
Figure 4: Attention visualization of the DMMGR model. The top and bottom maps are the attention weights of randomly selected samples from the test set that are correctly and incorrectly answered by the model, respectively. Different colors are used to represent the attention weights in the spatial-aware image graph and knowledge memory module, where deeper colors denote higher attention weights." Answer" represents the true answer and "Prediction" stands for the output of the model.(More examples can be found in Figure 5)

model in Table 4 has the regular key-value memory module whose keys are the representation of subject and relation, and values are the object representation of a knowledge triplet (Miller et al. 2016). The third model is the dynamic key-value memory module proposed in this paper. The results show that the accuracy of our proposed memory model surpasses the other two models by approximately 5%, which indicates that the dynamic key-value memory module has a beneficial effect on knowledge triplet reasoning. It is also rational to observe that a simple memory module slightly performs better than the common key-value memory module (Miller et al. 2016), since the latter can only infer the object of a triplet while questions in the KRVQR dataset could be related to the subject or relation of a triplet.

Memory model	Top-1 accuracy
Average embedding memory module	27.3
Key-value memory module	26.5
Proposed memory model	31.4

Table 4: Results in terms of top-1 accuracy of the DMMGR model obtained on the KRVQR dataset considering different types of memory modules.

Can External Knowledge Help the Model to Understand the Image Better?

Finally, we explore whether the retrieved knowledge triplets can help the model to better understand the image. In Ta-

ble 5, we compare the performance of the models with and without the use of the knowledge triplets in the reasoning module operating on the spatial-aware image graph. One can observe that the accuracy decreases by approximately 2% when the reasoning module has no access to the knowledge triplets, which supports our proposed knowledge-guided image graph reasoning module.

Model (DMMGR)	Accuracy
w/o knowledge-guided reasoning	29.9
full model	31.4

Table 5: The accuracy values obtained with and without the us of the knowledge triplets in the image graph reasoning module obtained on the KRVQR dataset.

Qualitative Analysis

To further study the working mechanism of the DMMGR model, we randomly select two samples from the test set that are correctly answered by our model, and present the corresponding attention weights on top of Figure 4 (more examples are found in Figure 5). Note, that to obtain a clear visualization map, we plot only the attention weights of some highly relevant knowledge triplets and objects in the image. The attention weights show that DMMGR not only correctly selects the most relevant knowledge triplets and addresses the correct elements, but also focuses on the related objects in the image. The bottom attention weight maps of Figure 4
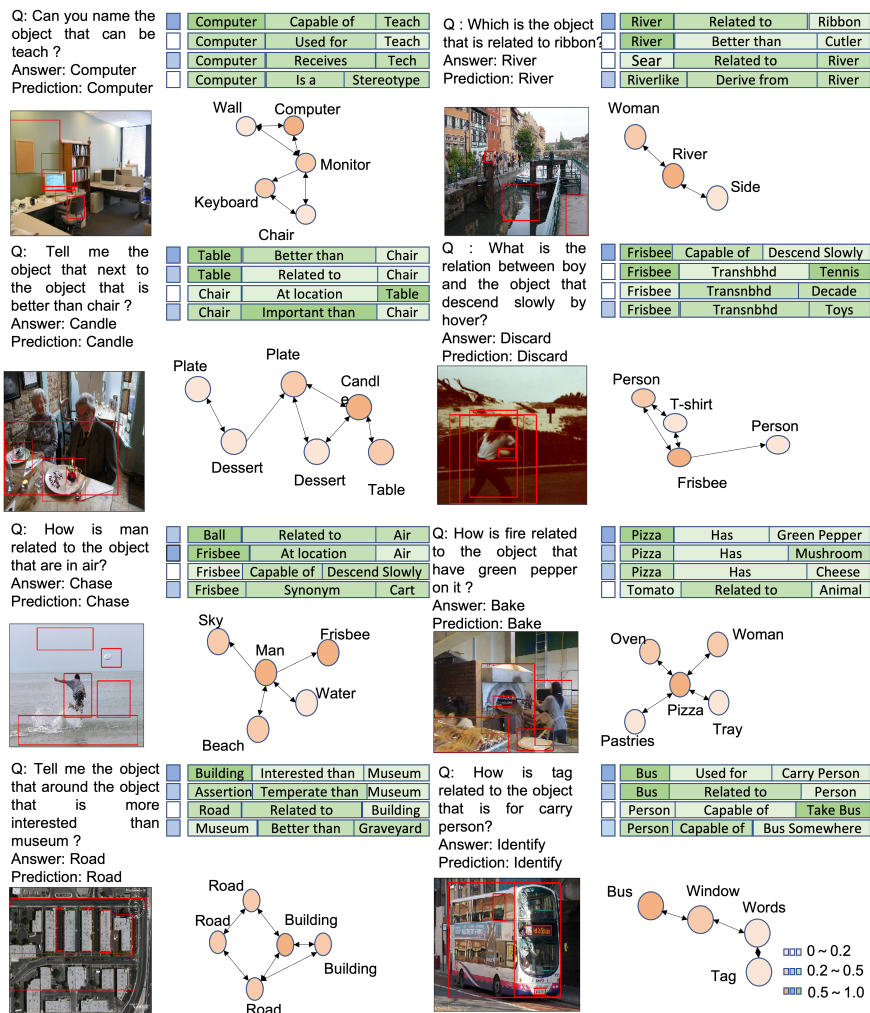
Figure 5: Attention visualization of the DMMGR model. All the samples are randomly selected from the test set of the KRVQR dataset. Different colors are used to represent the attention weights in the spatial-aware image graph and knowledge memory module, where deeper colors denotes higher attention weights." Answer" represents the true answer and "Prediction" stands for the output of the model.

are examples where the DMMGR model fails. We observe that DMMGR here fails when there is ambiguity present in the image. For example, there are more than two persons in the scene, although the model could find the most relevant information from the knowledge base, it focuses on the wrong person, which leads to an incorrect answer. Improving the model's ability of handling such confusion could be an interesting future work.

Conclusion

We have proposed a multi-step graph reasoning model that is enhanced by a novel dynamic memory, which iteratively performs explicit and implicit reasoning over a key-value triplet memory and a spatial-aware image graph, respectively, to infer the answer in a KVQA task. Our model achieves new state-of-the-art performance on both the KRVQR and FVQA datasets.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and 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*, 6077–6086.

Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 39–48.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zit-

nick, C. L.; and Parikh, D. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2425–2433.

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. 2007. DBpedia: A nucleus for a web of open data. In *The Semantic Web*, 722–735. Springer.

Ben-younes, H.; Cadene, R.; Cord, M.; and Thome, N. 2017. MUTAN: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Cao, Q.; Li, B.; Liang, X.; and Lin, L. 2019. Explainable high-order visual question reasoning: A new benchmark and knowledge-routed network. *arXiv preprint arXiv:1909.10128*.

Cao, Q.; Li, B.; Liang, X.; Wang, K.; and Lin, L. 2021. Knowledge-routed visual question reasoning: Challenges for deep representation embedding. *IEEE Transactions on Neural Networks and Learning Systems*.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Gu, J.; Zhao, H.; Lin, Z.; Li, S.; Cai, J.; and Ling, M. 2019. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1969–1978.

Hu, R.; Rohrbach, A.; Darrell, T.; and Saenko, K. 2019. Language-conditioned graph networks for relational reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10294–10303.

Hudson, D. A.; and Manning, C. D. 2019. GQA: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*, 3(8).

Jain, A.; Kothyari, M.; Kumar, V.; Jyothi, P.; Ramakrishnan, G.; and Chakrabarti, S. 2021. Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. *Advances in Neural Information Processing Systems*, 29: 289–297.

Ma, L.; Lu, Z.; and Li, H. 2016. Learning to answer questions from image using convolutional neural network. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*.

Marino, K.; Rastegari, M.; Farhadi, A.; and 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*, 3195–3204.

Miller, A.; Fisch, A.; Dodge, J.; Karimi, A.-H.; Bordes, A.; and Weston, J. 2016. Key-value memory networks for directly reading documents. *arXiv preprint arXiv:1606.03126*.

Narasimhan, M.; Lazebnik, S.; and Schwing, A. G. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. *arXiv preprint arXiv:1811.00538*.

Narasimhan, M.; and Schwing, A. G. 2018. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. *Proceedings of the European Conference on Computer Vision (ECCV)*, 451–468.

Norcliffe-Brown, W.; Vafeias, E.; and Parisot, S. 2018. Learning conditioned graph structures for interpretable visual question answering. *arXiv preprint arXiv:1806.07243*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32: 8026–8037.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; and Manning, C. D. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28: 91–99.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-first AAAI Conference on Artificial Intelligence*.

Sukhbaatar, S.; Szlam, A.; Weston, J.; and Fergus, R. 2015. End-to-end memory networks. *arXiv preprint arXiv:1503.08895*.

Tandon, N.; De Melo, G.; Suchanek, F.; and Weikum, G. 2014. Webchild: Harvesting and organizing commonsense knowledge from the Web. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, 523–532.

Teney, D.; Liu, L.; and van den Hengel, A. 2017. Graph-structured representations for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2017. Graph attention networks. *6th International Conference on Learning Representations*.

Wang, P.; Wu, Q.; Cao, J.; Shen, C.; Gao, L.; and van den Hengel, A. 2019. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1960–1968.

Wang, P.; Wu, Q.; Shen, C.; Dick, A.; and van den Hengel, A. 2017. FVQA: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10): 2413–2427.

Weston, J.; Chopra, S.; and Bordes, A. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

Xu, K.; Lai, Y.; Feng, Y.; and Wang, Z. 2019. Enhancing key-value memory neural networks for knowledge based question answering. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2937–2947.

Yang, Z.; He, X.; Gao, J.; Deng, L.; and Smola, A. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 21–29.

Yu, J.; Zhu, Z.; Wang, Y.; Zhang, W.; Hu, Y.; and Tan, J. 2020. Cross-modal knowledge reasoning for knowledge-based visual question answering. *Pattern Recognition*, 108: 107563.

Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6281–6290.

Yu, Z.; Yu, J.; Xiang, C.; Fan, J.; and Tao, D. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 29(12): 5947–5959.

Zareian, A.; Karaman, S.; and Chang, S.-F. 2020. Bridging knowledge graphs to generate scene graphs. In *Proceedings of the European Conference on Computer Vision*, 606–623. Springer.

Zhu, Z.; Yu, J.; Wang, Y.; Sun, Y.; Hu, Y.; and Wu, Q. 2020. Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 1097–1103. International Joint Conferences on Artificial Intelligence Organization.

Ziaeefard, M.; and Lécué, F. 2020. Towards knowledge-augmented visual question answering. In *Proceedings of the 28th International Conference on Computational Linguistics*, 1863–1873.