Pattern Recognition 108 (2020) 107563

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Cross-modal knowledge reasoning for knowledge-based visual question answering

Jing Yu[a,b,*], Zihao Zhu[a,b], Yujing Wang[c,d], Weifeng Zhang[e], Yue Hu[a,b], Jianlong Tan[a,b]

[a] Institute of Information Engineering, Chinese Academy of Sciences, China
[b] School of Cyber Security, University of Chinese Academy of Sciences, China
[c] Microsoft Research Asia, China
[d] Key Laboratory of Machine Perception, MOE, School of EECS, Peking University
[e] College of Mathematics, Physics and Information Engineering, Jiaxing University, China

ARTICLE INFO

Article history:
Received 12 March 2020
Revised 13 May 2020
Accepted 21 July 2020
Available online 22 July 2020

Keywords:
Cross-modal knowledge reasoning
Multimodal knowledge graphs
Compositional reasoning module
Knowledge-based visual question answering
Explainable reasoning

ABSTRACT

Knowledge-based Visual Question Answering (KVQA) requires external knowledge beyond the visible content to answer questions about an image. This ability is challenging but indispensable to achieve general VQA. One limitation of existing KVQA solutions is that they jointly embed all kinds of information without fine-grained selection, which introduces unexpected noises for reasoning the correct answer. How to capture the question-oriented and information-complementary evidence remains a key challenge to solve the problem. Inspired by the human cognition theory, in this paper, we depict an image by multiple knowledge graphs from the visual, semantic and factual views. Thereinto, the visual graph and semantic graph are regarded as image-conditioned instantiation of the factual graph. On top of these new representations, we re-formulate Knowledge-based Visual Question Answering as a recurrent reasoning process for obtaining complementary evidence from multimodal information. To this end, we decompose the model into a series of memory-based reasoning steps, each performed by a **G**raph-based **R**ead, **U**pdate, and **C**ontrol (**GRUC**) module that conducts parallel reasoning over both visual and semantic information. By stacking the modules multiple times, our model performs transitive reasoning and obtains question-oriented concept representations under the constrain of different modalities. Finally, we perform graph neural networks to infer the global-optimal answer by jointly considering all the concepts. We achieve a new state-of-the-art performance on three popular benchmark datasets, including FVQA, Visual7W-KB and OK-VQA, and demonstrate the effectiveness and interpretability of our model with extensive experiments. The source code is available at: https://github.com/astro-zihao/gruc

1. Introduction

Visual question answering (VQA) [1] is an attractive research direction aiming to jointly analyze multimodal content from images and natural language. Equipped with the capacities of grounding, reasoning and translating, a VQA agent is expected to answer a question in natural language based on an image. Recent works [2,3] have achieved great success in VQA tasks that are answerable by solely referring to the visible content. However, such kinds of models are incapable of answering questions which require external knowledge beyond the visible content. Considering the question in Fig. 1, the agent not only needs to visually localize 'red cylinder', but also to semantically recognize it as 'fire hydrant' and connects the knowledge that 'fire hydrant is used for firefighting'. Therefore, how to collect question-oriented and information-complementary evidence from visual, semantic and knowledge perspectives is essential to achieve general VQA.

To advocate research in this direction, [4] introduces a Knowledge-based Visual Question Answering (KVQA) task, named as 'Fact-based' VQA (FVQA), for answering questions by joint analysis of the image and the knowledge base of facts. The typical solutions for FVQA build a fact graph with fact triplets filtered by the visual concepts in the image and select one entity in the graph as the answer.

Existing works [4,5] parse the question as keywords and retrieve the supporting-entity only by keyword matching. This kind of approaches is vulnerable when the question does not exactly mention the visual concepts (*e.g.* synonyms and homographs) or the mentioned information is not captured in the fact graph (*e.g.* the visual attribute 'red' in Fig. 1 may be falsely omitted). To re-

* Corresponding author at: A3 Building, Minzhuang Road No. 89, Haidian District, Beijing, China.
E-mail address: yujing02@iie.ac.cn (J. Yu).

https://doi.org/10.1016/j.patcog.2020.107563
0031-3203/

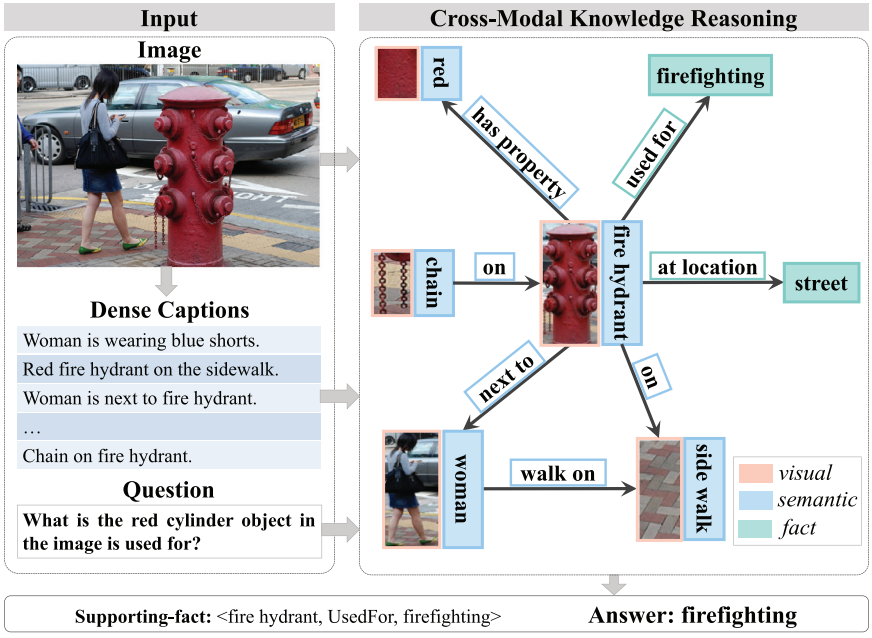

Fig. 1. An illustration of our motivation. We represent an image by graphs to associate visual, semantic and factual knowledge corresponding to the objects and relationships. Cross-modal knowledge reasoning is conducted on the graphs to infer the optimal answer.

solve these problems, [6] introduces visual information into the fact graph and infers the answer by implicit graph reasoning under the guidance of the question. However, they provide the whole visual information equally to each graph node by concatenation of the image, question and entity embeddings. Actually, only part of the visual content are relevant to the question and a certain entity. Moreover, the fact graph here is still homogeneous since each node is represented by a fixed form of image-question-entity embedding, which limits the model's flexibility of adaptively capturing evidence from different modalities. A model has to be selective by choosing relevant information and avoiding unexpected noise.

The recent proposed natural language understanding systems based on the cognition theory [7] are consistent in that our brain is capable to adaptively combine multimodal input for understanding and reasoning. As proposed in [7], the understanding system contains two essential parts, where the neocortical sub-system (the blue box in Fig. 2) is responsible for selectively integrating linguistic and non-linguistic input to understand the object and situation while the medial temporal lobe (MTL) sub-system (the red box in Fig. 2) aims to store and learn from the integrated embeddings of the neocortical sub-system states. This understanding system is universal to tackle a wide range of natural language problems requiring external knowledge in multimodal format. In this perspective, KVQA problems can also be solved by this system by considering external knowledge in both linguistic (text and knowledge graph) and non-linguistic (image) format, which is flexible to choose task-relevant and content-complementary information for answer prediction.

Motivated by the proposed structure of understanding system in McClelland et al. [7], we first introduce a novel scheme to depict an image by unifying the graph representation of different modalities, including the visual graph, semantic graph and fact graph. Specifically, the object appearance and their relationships are kept in the visual graph, the high-level abstraction is provided in the semantic graph and the corresponding factual knowledge is supported in the fact graph, which imitates distinct areas in neocortical sub-system that processes each input modality (the three blue ovals at bottom of neocortex). Graph representation is suitable for modeling the objects (entities) and their relationships of input and beneficial for connecting different modalities to constrain each other. Then we integrate cross-modal knowledge corresponding to the same concept (the blue oval at top of neocortex) by a series of memory-based reasoning steps. In order to select complementary knowledge from different modalities for integration, we propose a constraint satisfaction process in which the information in one modality influences the selection of information in another modality. To this end, we perform each reasoning step by a **G**raph-based **R**ead, **U**pdate, and **C**ontrol (**GRUC**) module that conducts parallel reasoning over both visual and semantic information: the control unit updates the control signal for extracting a knowledge vector from the knowledge graphs (visual and semantic); the read unit generates the knowledge vector from the knowledge graphs upon the constrain of the control signal; the update unit integrates the knowledge vector into the control signal as well as the knowledge graph for memory update. After multiple reasoning steps, we obtain complementary evidence from different modalities and fuse them adaptively to reason about the global-optimal answer via a graph neural network, which can be seen as the learning system in MLT. The main contributions can be summarized as follows:

(1) We novelly depict multimodal knowledge sources by multiple knowledge graphs from the visual, semantic and factual views, which unifies the representations of different modalities in graph domain and thus benefits for structure preserving and relational reasoning. Thereinto, introducing the semantic graph for high-level abstraction brings remarkable improvement in KVQA, which has been less studied in previous work.

(2) We propose a recurrent reasoning model that has three obvious novelties: First, it is a parallel reasoning model that applies modality-oriented controllers for reasoning over different modalities in a parallel mode, which can be easily extended to involve more modalities; Second, our model is designed for reasoning upon graph-structured multimodal data, aiming to consider the essential structural information in the reasoning process; Third, the basic reasoning module GRUC is a modular architecture consisting of Read, Update

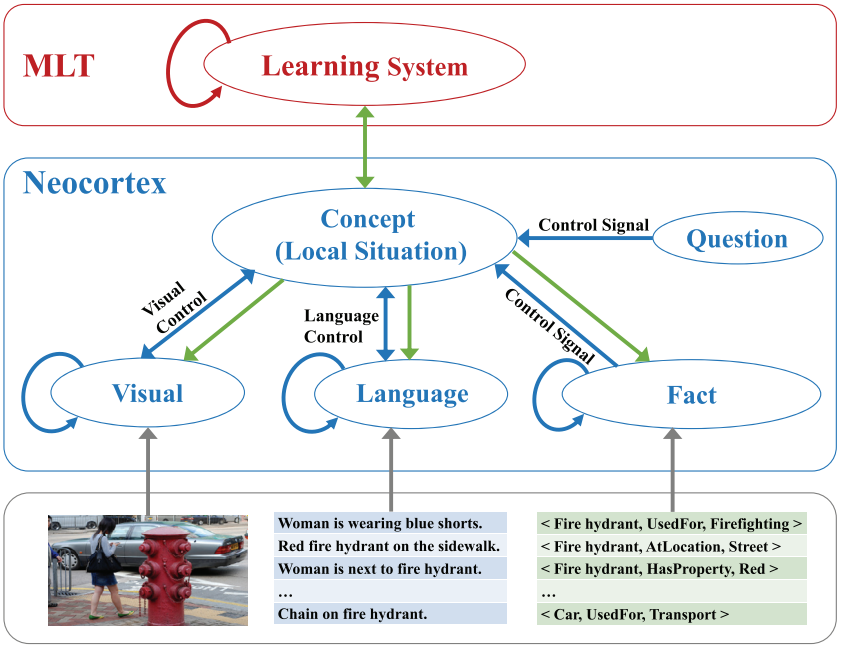

Fig. 2. The understanding system emulated by our model. The gray box contains information from multimodal sources. The blue box contains the neocortical sub-system, with each oval presenting an representation of a modal information. The blue arrows indicate learned connections that allow the representations to constrain each other. The red box contains the medial temporal lobe (MLT) sub-system, which stores and processes all the representations from the neocortical system. The red arrow represents self-connections that jointly consider all the representations for use. Green arrows connecting the red and blue ovals provide constraint between the two sub-systems. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and Control units, which supports more explicit and structured reasoning.

(3) The proposed model remarkably outperforms state-of-the-art approaches on three benchmark datasets, including FVQA, Visual7W-KB and OK-VQA, which demonstrates the feasibility and effectiveness of the proposed model. Through ablative studies, we prove how each of the proposed components contributes to the improvement.

(4) The proposed model has good interpretability. It automatically tells which concept (entity) and modality (visual, semantic or factual) have more contributions to answer the question through visualization of attention weights in GRUC and gate values in the fusion process. Meanwhile, the model can also reveal the knowledge selection mode from different modalities according to the complexity of the questions.

2. Related work

2.1. Visual question answering

The Visual Question Answering (VQA) task requires the agent to answer a question in natural language according to the visual content in an image, which demands for comprehending and reasoning about both visual and textual information. The typical solutions for VQA are based on the CNN-RNN architecture [8] that coarsely fuses the global visual and textual features as clues to predict the answer. For better combining textual information with visual information, bilinear pooling approaches [9,10] have been proposed to fuse multimodal features in fine-grained mode. However, the above approaches leverage all the information in the image and question, which may introduce redundant or noisy information to the prediction stage. To alleviate this problem, various attention mechanisms have been exploited in VQA tasks [11,12] to highlight visual objects that are relevant to the question. Anderson et al. [13] introduced a bottom-up and top-down attention mechanism to learn the attention on candidate objects rather than spatial grids. However, they treat objects in an image independently and ignore their informative relationships.

Humans' ability of combinatorial generalization highly depends on the mechanisms of reasoning over relationships. Consistent with such idea, there is an emerging trend to depict objects and visual relationships in an image by graph structure to support reasoning in VQA. One kind of approaches performs one-step relational reasoning to infer the answer. Santoro et al. [14] proposed a Relation Network (RN) to model all the implicit relationships among objects in the image by multi-layer perceptrons (MLPs). Then, the relationships are summed and fed into other MLPs to predict the answer. This approach brings high computational cost and makes it hard to perform multi-step reasoning. Norcliffe-Brown et al. [15] refined the relationships by a ranking strategy and lowered the computational complexity. However, a large number of compound questions require multi-step reasoning. To this end, Wu et al. [16] proposed multi-step attention to reason over both original objects and new compound objects and infer the answer progressively. Additionally, Hudson and Manning [17] transformed both the visual and textual moldalities into concept-based graph representations and performed sequential reasoning over the graph by the neural state machine. Furthermore, Jiang et al. [18] exploited semantic captions to further enrich the graph-based representations for multi-step reasoning. Reasoning approaches in the above work are always on visual and textual features, which cannot be extended to involve external knowledge. To go one step further, our model pays attention to not only original input features but also external knowledge during progressive reasoning.

2.2. Incorporating external knowledge in VQA

Human easily combine visual observation with external knowledge for answering questions, which remains challenging for algorithms. To bridge this discrepancy, Wang et al. [4] introduced a Fact-based VQA (FVQA) task, which additionally provides a knowledge base of facts and associates each question with a supporting-

fact. Recent works based on FVQA generally select one entity from fact graph as the answer and falls into two categories: query-mapping based methods and learning based methods. On the one hand, Wang et al. [5] reduced the question to one of the available query templates and this limits the types of questions that can be asked. Wang et al. [4] automatically classified and mapped the question to a query which does not suffer this constraint. Among both methods, however, visual information is used to extract facts but not introduced during the reasoning process. On the other hand, Narasimhan and Schwing [19] learned a similarity score between the representations of fact and image-question pair. Narasimhan et al. [6] applied graph convolutional networks on the fact graph where each node is represented by the fixed form of image-question-entity embedding. However, the visual information is wholly provided which may introduce redundant information for reasoning the answer. The same problem also exists in Li et al. [20], although they leveraged dynamic memory network instead of graph convolutional netowrk to incorporate the external knowledge. Recent work [21] proposed a new knowledge-based task OK-VQA and introduced a retrieval-based model to extract the correct answer from Wikipedia. Different from previous work, in this paper, we decipt an image by multimodal kwnoledge graphs and perform cross-modal reasoning via a memory-based recurrent network to capture complementary evidence from different modalities.

2.3. Graph neural networks

The core module GRUC in our proposed model is a novel graph-based neural network. In this subsection, we briefly review related Graph Neural Networks (GNNs) and highlight differences between previous work and ours. Approaches based on GNNs [22] repeatedly perform a message passing process over the graph by aggregating and updating information between nodes. Relying on spectral graph theory, Kipf and Welling [23] exploited simplified Chebyshev polynomials to construct localized polynomial filters for graph convolution in graph convolutional networks (GCN). Attention mechanisms have been introduced in Veličković et al. [24] to learn the weights over edges for convolution operations. Lanczos-based method [25] has been explored for graph convolution for the purpose of acceleration. Our model is closely related to the Gated Graph Sequence Neural Networks (GGS-NN) [26] which updates GNNs by adding gated recurrent unit. Different from GGS-NN that both convolution operation and recurrent propagation are performed in the same graph, our GRUC module aggregates information from external knowledge graph for node updating and recurrent propagation in another fact graph. Our model is also related to the heterogeneous graph neural networks since the model is reasoning over multimodal graphs. Schlichtkrull et al. [27] generalized graph convolutional network to handle different relationships between entities in a knowledge base, where edges with distinct relationships are encoded independently. Wang et al. [28] proposed heterogeneous graph attention networks with dual-level attention mechanism. All the above approaches model different types of nodes and edges in an unified graph. In contrast, the heterogeneous graph in this work contains multiple layers of subgraphs and each layer consists of nodes and edges coming from different modalities. For this specific constrain, we propose the parallel reasoning model that applies modality-oriented controllers for reasoning over different modalities in a parallel mode.

3. Methodology

Given an image I and a question Q, the task aims to predict an answer A by leveraging the external knowledge. In this work, we focus on external knowledge in the form of knowledge graph, which consists of a set of triplet facts, *i.e.* $< e_1, r, e_2 >$, where e_1 is a visual concept in the image, e_2 is an attribute or phrase and r represents the relationship between e_1 and e_2. The key is to choose a correct concept, *i.e.* either e_1 or e_2, from the supporting fact as the predicted answer.

The proposed model mainly contains four parts: (1) Firstly, *multimodal Knowledge Graph Construction* (Section 3.1) represents knowledge from different modalities by different knowledge graphs, including the visual graph, semantic graph and fact graph, imitating distinct brain areas that represent each input modality; (2) Then, *Intra-Modal Knowledge Selection* (Section 3.2) selects question-oriented knowledge from each modality of knowledge graph by intra-modal graph convolution; (3) Afterwards, *Cross-Modal Knowledge Reasoning* (Section 3.3) performed by the GRUC Network iteratively gathers complementary evidence from the visual and semantic knowledge graphs under the guidance of the question and the facts. In the end of reasoning steps, we fuse the evidence from three modalities to obtain the representation of each concept. (4) Finally, *Global Assessment and Answer Prediction* (Section 3.4) aims to jointly analyze all the concepts via graph convolutional networks and predict the optimal answer by a binary classifier. Fig. 3 gives detailed illustration of our model.

3.1. Multimodal graph construction

3.1.1. Visual graph construction

Since most of the questions in KVQA grounded in the visual objects and their relationships, we construct a fully-connected visual graph to represent such evidence at appearance level. Given an image I, we use Faster-RCNN [29] to identify a set of objects $\mathcal{O} = \{o_i\}_{i=1}^{K}$ ($K = 36$), where each object o_i is associated with a visual feature vector $\boldsymbol{v}_i \in \mathbb{R}^{d_v}$ ($d_v = 2048$), a bounding-box feature vector $\boldsymbol{b}_i \in \mathbb{R}^{d_b}$ ($d_b = 4$) and a corresponding label. Specifically, $\boldsymbol{b}_i = [x_i, y_i, w_i, h_i]$, where (x_i, y_i), h_i and w_i respectively denote the coordinate of the top-left corner, the height and width of the bounding box. We construct a visual graph $\mathcal{G}^V = (\mathcal{V}^V, \mathcal{E}^V)$ over $\mathcal{O}$, where $\mathcal{V}^V = \{v_i^V\}_{i=1}^{K}$ is the node set and each node v_i^V corresponds to a detected object o_i. The feature of node v_i^V is represented by $\boldsymbol{v}_i^V$. Each edge $e_{ij}^V \in \mathcal{E}^V$ denotes the relative spatial relationships between two objects. We encode the edge feature by a 5-dimensional vector, *i.e.* $\boldsymbol{r}_{ij}^V = [\frac{x_j - x_i}{w_i}, \frac{y_j - y_i}{h_i}, \frac{w_j}{w_i}, \frac{h_j}{h_i}, \frac{w_j h_j}{w_i h_i}]$.

3.1.2. Semantic graph construction

In addition to visual information, high-level abstraction of the objects and relationships by natural language provides essential semantic information. Such abstraction is indispensable to associate the visual objects in the image with the concepts mentioned in both questions and facts. In our work, we leverage dense captions [30] to extract a set of local-level semantics in an image, ranging from the properties of a single object (color, shape, emotion, *etc.*) to the relationships between objects (action, spatial positions, comparison, *etc.*). We decipt an image by D dense captions, denoted as $Z = \{z_i\}_{i=1}^{D}$, where z_i is a natural language description about a local region in the image. Instead of using monolithic embeddings to represent the captions, we exploit modeling them by a graph representation, denoted as $\mathcal{G}^S = (\mathcal{V}^S, \mathcal{E}^S)$, which is constructed by a semantic graph parsing model [31]. The node $v_i^S \in \mathcal{V}^S$ represents the name or attribute of an object extracted from the captions while the edge $e_{ij}^S \in \mathcal{E}^S$ represents the relationship between v_i^S and v_j^S. We use the averaged GloVe embeddings to represent v_i^S and e_{ij}^S, denoted as $\boldsymbol{v}_i^S$ and $\boldsymbol{r}_{ij}^S$, respectively. The graph representation retains the relational information among concepts and unifies the representations in graph domain, which is better for explicit reasoning across modalities.

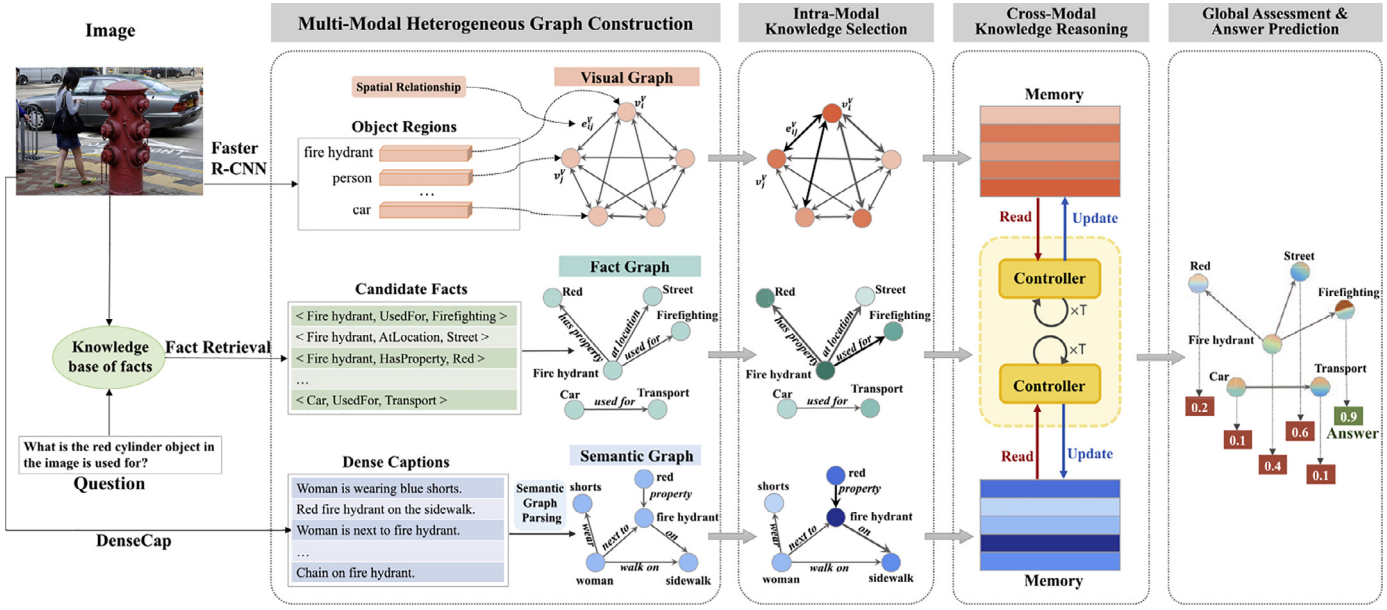

Fig. 3. An overview of our model. The model contains four parts: multimodal heterogeneous graph construction, intra-Modal knowlegdge selection, cross-modal knowledge reasoning, and global assessment and answer prediction.

3.1.3. Fact graph construction

To find the optimal supporting-fact, we first retrieve relevant candidate facts from knowledge base of facts following a score-based approach [6]. We compute the cosine similarity of the GloVe embeddings of every word in the fact with the words in the question and the words of visual concepts detected in the image. Then we average these values to assign a similarity score to the fact. The facts are sorted based on the similarity and the 100 highest scoring facts are retained, denoted as f_{100}. A relation type classifier is trained additionally to further filter the retrieved facts. Specifically, we feed the last hidden state of LSTM to an MLP layer to predict the relation type $\hat{r}_i$ of a question. We retain the facts among f_{100} only if their relationships agree with $\hat{r}_i$, *i.e.* $f_{rel} = f \in f_{100} : r(f) \in \{\hat{r}_i\}$ ($\{\hat{r}_i\}$ contains top-3 predicted relationships in experiments). Then a fact graph $\mathcal{G}^F = (\mathcal{V}^F, \mathcal{E}^F)$ is built upon f_{rel} as the candidate facts can be naturally organized as graphical structure. Each node $v_i^F \in \mathcal{V}^F$ denotes an entity in f_{rel} and is represented by GloVe embedding of the entity, denoted as $\boldsymbol{v}_i^F$. Each edge $e_{ij}^F \in \mathcal{E}^F$ denotes the relationship between v_i^F and v_j^F and is represented by GloVe embedding $\boldsymbol{r}_{ij}$. The topological structure among facts can be effectively exploited by jointly considering all the entities in the fact graph.

3.2. Intra-modal knowledge selection

Since each layer of graphs contains modality-specific knowledge relevant to the question, we first select valuable evidence independently from the visual graph, semantic graph and fact graph by *Visual-to-Visual Convolution*, *Semantic-to-Semantic Convolution* and *Fact-to-Fact Convolution*, respectively. These three convolutions share the common operations but differ in their node and edge representations corresponding to the graph layers. Thus we omit the superscript of node representation $\boldsymbol{v}$ and edge representation $\boldsymbol{r}$ in the rest of this section. We first perform attention operations to highlight the nodes and edges that are most relevant to the question q and consequently update node representations via intra-modal graph convolution. This process mainly consists of the following three steps:

Question-guided node attention. We first evaluate the relevance of each node corresponding to the question by attention mechanism. The attention weight for v_i is computed as:

$$\alpha_i = \text{softmax}(\boldsymbol{w}_a^T \tanh(\mathbf{W}_1 \boldsymbol{v}_i + \mathbf{W}_2 \boldsymbol{q})) \tag{1}$$

where $\mathbf{W}_1, \mathbf{W}_2$ and $\boldsymbol{w}_a$ (as well as $\mathbf{W}_3$,..., $\mathbf{W}_{12}$, $\boldsymbol{w}_b$, $\boldsymbol{w}_c$ mentioned below) are learned parameters. $\boldsymbol{q}$ is question embedding encoded by the last hidden state of LSTM.

Question-guided edge attention. Under the guidance of question, we then evaluate the importance of edge e_{ji} constrained by the neighbor node v_j regarding to v_i as follows:

$$\beta_{ji} = \text{softmax}(\boldsymbol{w}_b^T \tanh(\mathbf{W}_3 \boldsymbol{v'}_j + \mathbf{W}_4 \boldsymbol{q'})) \tag{2}$$

where $\boldsymbol{v'}_j = \mathbf{W}_5[\boldsymbol{v}_j, \boldsymbol{r}_{ji}]$, $\boldsymbol{q'} = \mathbf{W}_6[\boldsymbol{v}_i, \boldsymbol{q}]$ and $[\,\cdot\,,\,\cdot\,]$ denotes concatenation operation.

Intra-modal graph convolution. Given the node and edge attention weights learned in Eqs. (1) and (2), the node representations of each layer of graphs are updated following the message-passing framework [32]. We gather the neighborhood information and update the representation of v_i as follows:

$$\boldsymbol{m}_i = \sum_{j \in \mathcal{N}_i} \beta_{ji} \boldsymbol{v'}_j \tag{3}$$

$$\hat{\boldsymbol{v}}_i = \text{ReLU}(\mathbf{W}_7[\boldsymbol{m}_i, \alpha_i \boldsymbol{v}_i]) \tag{4}$$

where $\mathcal{N}_i$ is the neighborhood set of node v_i. We conduct the above intra-modal knowledge selection on $\mathcal{G}^V$, $\mathcal{G}^S$ and $\mathcal{G}^F$ independently and obtain the updated node representations, denoted as $\{\hat{\boldsymbol{v}}_i^V\}_{i=1}^{\mathcal{N}^V}$, $\{\hat{\boldsymbol{v}}_i^S\}_{i=1}^{\mathcal{N}^S}$ and $\{\hat{\boldsymbol{v}}_i^F\}_{i=1}^{\mathcal{N}^F}$ accordingly.

3.3. Cross-modal knowledge reasoning

The key of cross-modal knowledge graph reasoning module is to reason on the fact graph by jointly considering all candidate facts. However, the entities in the fact graph provide insufficient knowledge to reason about the globally optimal answer which need to be complemented with correlated knowledge from other modalities. The process is performed by our proposed GRUC Network, a memory-based reasoning architecture by sequencing a recurrent Graph-based Read, Update and Control (GRUC) module. After multi-step reasoning, we fuse the multimodal knowledge for each entity and achieve more comprehensive understanding of an

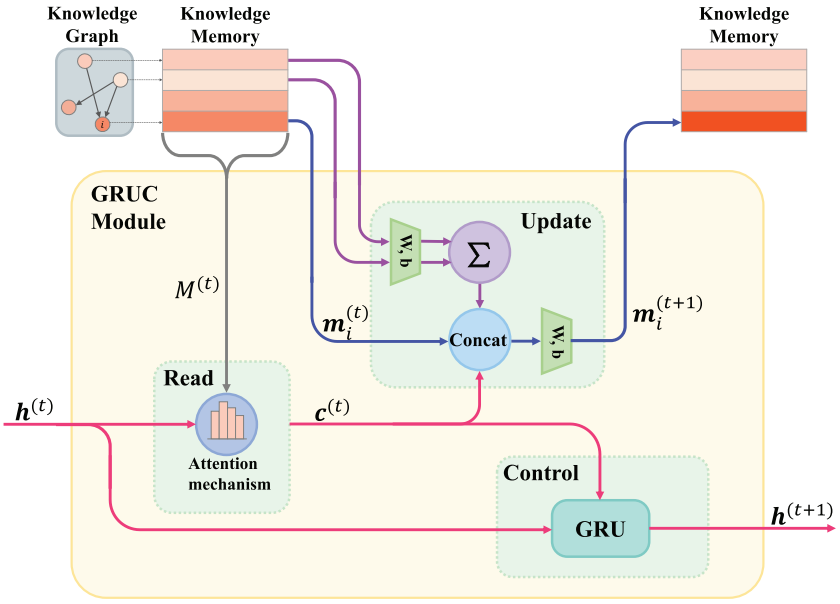

Fig. 4. The GRUC module architecture.

entity, which we rename it as 'concept' afterwards. The GRUC Network contains two components: a recurrent GRUC module and a multimodal feature fusion module, as introduced below.

3.3.1. The recurrent GRUC module

The GRUC module aims to gather question-oriented knowledge from different modalities corresponding to the same concept. However, it's non-trivial to align the knowledge from different modalities to the same concept since there rarely exists explicit one-to-one mappings across modalities. Therefore, we propose to gather concept-relevant knowledge in an implicit way. Since the answer comes from one entity in the fact graph in this task, we regard each entity in the fact graph as a concept and gather complementary knowledge from the visual graph and the semantic graph to this concept in the fact graph in the GRUC module. In this way, the recurrent GRUC modules can be performed over all the concepts in the fact graph in parallel and obtaining the relevant visual knowledge, semantic knowledge and fact knowledge for each concept as the output.

Fig. 5 shows the GRUC network operated on the concept (i.e. fact node) $v_i^F \in \mathcal{V}^F$. The recurrent GRUC modules (yellow boxes) iteratively gather knowledge from the visual knowledge memory (red box) and the semantic memory (blue box) in a parallel way. For each step t ($t = 1, 2, \ldots, T$) in the reasoning process, the tth GRUC module maintains two hidden states: visual control state $\boldsymbol{h}_v^{(t)}$ and semantic control state $\boldsymbol{h}_s^{(t)}$, initialized by the question q, v_i^F and its neighborhood nodes (gray and green boxes) to learn $\boldsymbol{h}_v^{(1)}$ and $\boldsymbol{h}_s^{(1)}$, respectively. Since the reasoning processes for the visual part and the semantic part share thee common operations but differ in the graph representations, we omit the subscript of hidden state representation $\boldsymbol{h}^{(t)}$ in the rest of this section. Fig. 4 shows the architecture of the GRUC module, which consists of the Control Unit, Read Unit and Update Unit.

The control unit. The control unit determines the *control state* $\boldsymbol{h}^{(t)}$ that guides the module to adaptively select complementary knowledge from the knowledge memory for obtaining a more comprehensive concept representation. At the first reasoning step, the control unit is initialized by fusing the question representation $\boldsymbol{q}$, the entity representation $\hat{\boldsymbol{v}}_i^F$ and its neighborhood information as:

$$\boldsymbol{h}^{(1)} = \mathbf{W}_8[\boldsymbol{q}, \hat{\boldsymbol{v}}_i^F, \boldsymbol{c}_i^F] \tag{5}$$

$$\boldsymbol{c}_i^F = \sum_{j \in \mathcal{N}_i} \hat{\boldsymbol{v}}_j^F \tag{6}$$

where $\mathcal{N}_i$ represents a set of 1-hop neighboring nodes regarding the entity v_i^F. In the tth reasoning step, the control state will be updated with the contextual vector $\boldsymbol{c}^{(t)}$ (will be introduced in the Read Unit) extracted from the knowledge memory and previous hidden state $\boldsymbol{h}^{(t)}$ via Gated Recurrent Unit (GRU) [33], the update operation is defined as follows:

$$\boldsymbol{h}^{(t+1)} = GRU(\boldsymbol{h}^{(t)}, \boldsymbol{c}^{(t)}) \tag{7}$$

Then updated control state $\boldsymbol{h}^{(t+1)}$ is used to control the reasoning process in the next step.

The read unit. In the tth reasoning step, the read unit gathers the required knowledge $\boldsymbol{c}^{(t)}$ in the knowledge base (visual graph or semantic graph) under the guidance of the control state $\boldsymbol{h}^{(t)}$ and previous knowledge memory $\boldsymbol{M}^{(t)} = \{\boldsymbol{m}_1^{(t)}, \boldsymbol{m}_2^{(t)}, \ldots, \boldsymbol{m}_N^{(t)}\}$ (N is the number of memory entries). Specifically, $\boldsymbol{M}^{(t)}$ is a graph-structured memory, where each entry represents the node in the knowledge graph ($\mathcal{G}^V$ or $\mathcal{G}^S$) and the relationships between entries are the corresponding edges in the knowledge graph. The initial representation of each entry $\boldsymbol{m}_j^{(1)}$ is the corresponding node representation obtained after intra-modal knowledge selection, *i.e.* $\boldsymbol{m}_j^{(1)} = \hat{\boldsymbol{v}}_j$.

We compute the required knowledge $\boldsymbol{c}^{(t)}$ via an attention component. The attention value $\gamma_j^{(t)}$ for the memory entry $\boldsymbol{m}_j^{(t)}$ is calculated under the guidance of the control state $\boldsymbol{h}^{(t)}$ as:

$$\boldsymbol{a}_j^{(t)} = \tanh(\mathbf{W}_9 \boldsymbol{h}^{(t)} + \mathbf{W}_{10} \boldsymbol{m}_j^{(t)}) \tag{8}$$

$$\gamma_j^{(t)} = \text{softmax}(\boldsymbol{w}_c^T \boldsymbol{a}_j^{(t)}) \tag{9}$$

Then we generate the required knowledge $\boldsymbol{c}^{(t)}$ by weighting over all the memory entries defined as:

$$\boldsymbol{c}^{(t)} = \sum_{j=1}^{N} \gamma_j^{(t)} \boldsymbol{m}_j^{(t)} \tag{10}$$

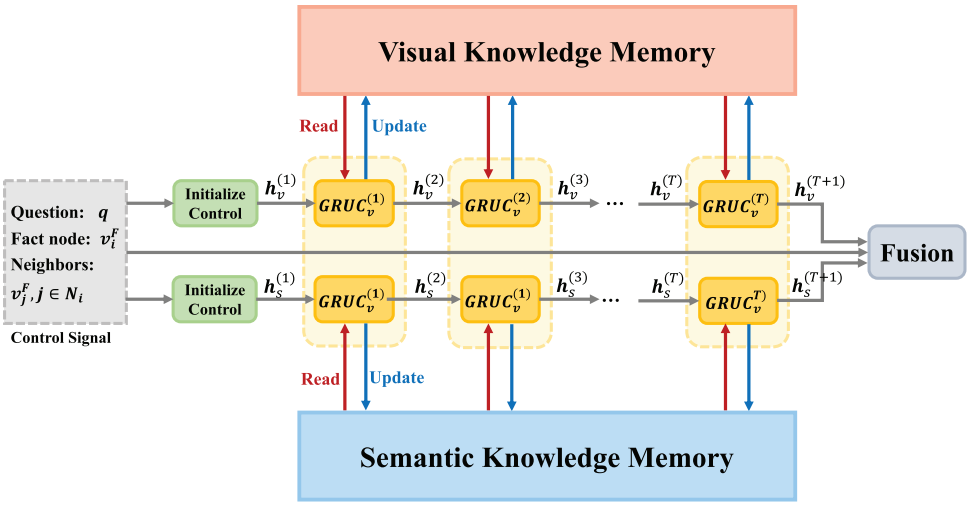

Fig. 5. An illustration of the GRUC network.

The function of $\boldsymbol{c}^{(t)}$ is divided into two folds: On the one hand, it is used to update the control unit and generate the control state $\boldsymbol{h}^{(t+1)}$ for the next reasoning step; On the other hand, $\boldsymbol{c}^{(t)}$ records the required knowledge retrieved from the knowledge graph during the transitive reasoning process and will be regarded as the knowledge representation of corresponding modality for the final multimodal feature fusion.

The update unit. Unlike the static knowledge representations, our proposed knowledge memory will be updated adaptively during each reasoning step via the Update Unit. This mechanism aims to enable the knowledge memory to remember what knowledge has been used and update the memory accordingly. Another difference compared with traditional knowledge representations lies in that our proposed knowledge memory is graph-structured. Updating each memory entry will stimulate its neighboring entries and transmit their information in the update operation. Formally, in tth reasoning step, each memory entry is updated based on its previous memory state, its neighborhood's previous state and the current control state as:

$$m_j^{(t+1)} = \mathbf{W}_{11}[\boldsymbol{m}_j^{(t)}, \boldsymbol{c}_j^{nei}, \boldsymbol{h}^{(t)}] \tag{11}$$

$$\boldsymbol{c}_j^{nei} = \sum_{k \in \mathcal{N}_j} \mathbf{W}_{12}[\boldsymbol{m}_k^{(t)}, \boldsymbol{r}_{jk}] \tag{12}$$

where $\mathcal{N}_i$ represents a set of 1-hop neighboring nodes regarding the memory entity m_j and $\boldsymbol{c}_j^{nei}$ is the contextual memory representation. Then the updated memory is served as the new knowledge memory used in the next reasoning step.

3.3.2. Multimodal feature fusion module

After T reasoning steps, the model collects concept-relevant knowledge for concept v_i^F from the visual graph and the semantic graph independently and generate the corresponding knowledge representations denoted as $\boldsymbol{h}_{V_i}^{(T+1)}$ and $\boldsymbol{h}_{S_i}^{(T+1)}$ respectively. Then we fuse the complementary knowledge from the three modalities to form the final concept representation $\boldsymbol{v}_i^F$ via gate mechanism as:

$$\boldsymbol{gate}_i = \sigma\left(\mathbf{W}_{10}[\boldsymbol{h}_{V_i}^{(T+1)}, \boldsymbol{h}_{S_i}^{(T+1)}, \hat{\boldsymbol{v}}_i^F]\right) \tag{13}$$

$$\widetilde{\boldsymbol{v}}_i^F = \mathbf{W}_{11}\left(\boldsymbol{gate}_i \circ [\boldsymbol{h}_{V_i}^{(T+1)}, \boldsymbol{h}_{S_i}^{(T+1)}, \hat{\boldsymbol{v}}_i^F]\right) \tag{14}$$

where σ is sigmoid function and '$\circ$' is element-wise product.

3.4. Global assessment and answer prediction

All the concepts $\{\widetilde{\boldsymbol{v}}_i^F\}_{i=1}^{\mathcal{N}^F}$ are fed into a graph neural network [34] to globally compare with each other, which imitates the MLT in our understanding system. The output embedding of each concept in GNN is passed to a binary classifier to predict its probability as the answer, *i.e.* $\hat{y}_i = p_\theta([\widetilde{\boldsymbol{v}}_i^F, \boldsymbol{q}])$. Since there is one entity annotated as the ground-truth answer and the rest entities are all served as negative answers in each training sample, it is necessary to use weighted binary cross-entropy loss to deal with the imbalanced training data as:

$$l_n = -\sum_{i \in \mathcal{N}^F} \left[a \cdot y_i \ln \hat{y}_i + b \cdot (1 - y_i) \ln(1 - \hat{y}_i)\right] \tag{15}$$

where y_i is the ground truth label for v_i^F and a, b represent loss function weights for positive and negative samples respectively. The entity corresponding to the concept with the largest probability is selected as the final answer.

4. Experiments

4.1. Datasets and evaluation metrics

FVQA: The FVQA dataset [4] consists of 2190 images, 5286 questions and a knowledge base of 193,449 facts. The knowledge base is constructed by extracting the top visual concepts from all the images in the dataset and querying those concepts from three knowledge bases, including DBPedia [35], ConceptNet [36] and WebChild [37]. For each image-question pair in the dataset, the task aims to choose an entity in a supporting fact from the knowledge base as the answer by jointly considering the given question and image.

Visual7W + KB: The Visual7W dataset [38] is built based on a subset of images from Visual Genome [39], which includes questions in terms of (what, where, when, who, why, which and how) along with the corresponding answers in a multi-choice format. However, most of questions of Visual7W solely base on the image content which don't require external knowledge. Furthermore, Li et al. [20] generated a collection of knowledge-based questions based on the test images in Visual7W by filling a set of question-answer templates that need to reason on both visual content and external knowledge. We denoted this dataset as Visual7W + KB in our paper. In general, Visual7W + KB consists of 16,850 open-domain question-answer pairs based on 8425 images in Visual7W test split. Different from FVQA, Visual7W + KB uses ConceptNet to

guide the question generation but doesn't provide a task-specific knowledge base. In our work, we also leverage ConceptNet to retrieve the supporting knowledge and select one entity as the predicted answer.

OK-VQA: [21] proposed the Outside Knowledge VQA (OK-VQA) dataset, which is the largest KVQA dataset at present. Different from existing KVQA datasets, the questions in OK-VQA are manually generated by MTurk workers, which are not derived from specific knowledge bases. Therefore, it requires the model to retrieve supporting knowledge from open-domain resources, which is much closer to the general VQA but more challenging for existing models. OK-VQA contains 14,031 images which are randomly collected from MSCOCO dataset [40], using the original 80k-40k training and validation splits as train and test splits. OK-VQA contains 14,055 questions covering a variety of knowledge categories such as science & technology, history, and sports. In our work, we leverage ConceptNet to retrieve the supporting knowledge and select one entity as the predicted answer.

Evaluation metrics: We follow the metrics in Wang et al. [4] to evaluate the question answering performance. The top-1 and top-3 accuracy is calculated for each model. The averaged accuracy of 5 test splits is reported as the overall accuracy.

4.2. Implementation details

For the question representation, each question is tokenized and each word is embedded using 300-dimensional GloVe word embeddings [41]. The maximum sentence length of question is set to 20 and questions shorter than 20 words are padded with zero vectors. The sequence of embedded words is then fed into LSTM and the dimension of the hidden layer in LSTM is set to 512. For constructing the semantic graph, we first generate dense captions with DenseCap [30]. Since some captions with low confidence are likely to introduce unexpected noise and too many captions will decrease the computation efficiency, we select top-12 dense captions with highest scores to eliminate unexpected noise caused by low confidence captions. For the image representation, we extract 2048-dimensional object detection features, 4-dimensional spatial features, object labels with known bounding boxes from pre-trained Faster R-CNN [29] model in conjunction with ResNet-101 [42]. The number of detected objects is fixed to 36. The Faster R-CNN model is trained over 1600 selected object classes and 400 attribute classes, in a similar way as the bottom-up attention model [13].

We set $a = 0.7$ for positive samples and $b = 0.3$ for negative samples in the binary cross-entropy loss function. Our model is trained by Adam optimizer with 10 epochs, where the mini-batch size is 16 and the dropout ratio is 0.5. For the strategy of learning rate, we first apply warm up strategy for 2 epoches with initial learning rate 1×10^{-3} and warm-up factor 0.2. Then we adopt cosine annealing learning strategy with initial learning rate $\eta_{\max} = 1 \times 10^{-3}$ and termination learning rate $\eta_{\min} = 3.6 \times 10^{-4}$ for the rest epoches.

4.3. Comparison with state-of-the-art methods

4.3.1. Experimental results on FVQA

We compare our GRUC model with state-of-the-art models on FVQA dataset in Table 1. The baseline models can be classified into three sets, including CNN-RNN based approaches, semantic parsing based approaches and learning-based approaches. The CNN-RNN based approaches [4] include LSTM-Q + Image + Pre-VQA and Hie-Q + Image + Pre-VQA. The semantic parsing based approaches [4] include FVQA (top-3-QQmaping) and FVQA (Ensemble). The learning based approaches include Straight to the Facts (STTF) [19], Out of the Box (OB) [6], and Reading Comprehension based approach [43].

Table 1
State-of-the-art comparison on FVQA dataset.

Method	Overall accuracy	
	top-1	top-3
LSTM-Q + Image + Pre-VQA [4]	24.98	40.40
Hie-Q + Image + Pre-VQA [4]	43.14	59.44
FVQA (top-3-QQmaping) [4]	56.91	64.65
FVQA (Ensemble) [4]	58.76	–
Straight to the Facts (STTF) [19]	62.20	75.60
Reading Comprehension [43]	62.96	70.08
Out of the Box (OB) [6]	69.35	80.25
Human [4]	77.99	–
GRUC (ours)	**79.63**	**91.20**

Table 2
State-of-the-art comparison on Visual7W + KB dataset.

Method	Overall accuracy	
	top-1	top-3
KDMN-NoKnowledge [20]	45.1	–
KDMN-NoMemory [20]	51.9	–
KDMN [20]	57.9	–
KDMN-Ensemble [20]	60.9	–
Out of the Box (OB) [6]	57.32	71.61
GRUC (ours)	**69.03**	**88.12**

Our model consistently outperforms all the approaches on all the metrics and achieves remarkable 10.28% boost on top-1 accuracy and 10.95% boost on top-3 accuracy compared with the state-of-the-art model OB [6]. The model OB is most relevant to GRUC in that it leverages graph convolutional networks to jointly assess all the entities in the fact graph. However, it introduces the global image features equally to all the entities without selection. By collecting question-oriented visual and semantic information via the memory-based recurrent reasoning network, our model gains remarkable improvement. It's worth to note that our model outperforms Human performance by 1.64% on top-1 accuracy. To some extent, this results prove the effectiveness of the proposed understanding system in McClelland et al. [7] since our model is designed by emulating the structure of the proposed system. In other words, the model contains distinct modules that represent each input modality and computes the representation of concepts through a mutual constraint to combine linguistic and non-linguistic inputs. Emulating this architecture in model design could contribute to achieving human-level understanding ability.

4.3.2. Experimental results on Visual7W-KB

The comparison of state-of-the-art models on Visual7W-KB dataset is shown in Table 2. The compared baselines contains two sets, i.e. memory-based approaches and a graph-based approach. The memory-based approaches [20] include KDMN-NoKnowledge (w/o external knowledge), KDMN-NoMemory (attention-based knowledge incorporation), KDMN (dynamic memory network based knowledge incorporation) and KDMN-Ensemble (several KDMN models based ensemble model). We also test the performance of Out of the Box (OB) [6] on Visual7W-KB and report the results in Table 2.

As consistent with the results on FVQA, we achieve a significant improvement (8.13% on top-1 accuracy and 16.51% on top-3 accuracy) over state-of-the-art models. Note that our proposed GRUC network is an single-model, which outperforms the existing ensembled model [20]. We believe that the performance can be further improved if the technique of ensemble is involved in our model.

Table 3
State-of-the-art comparison on OK-VQA dataset. We show the results for the full OK-VQA dataset and for each knowledge category (top-1 accuracy): Vehicles and Transportation (VT); Brands, Companies and Products (BCP); Objects, Material and Clothing (OMC); Sports and Recreation (SR); Cooking and Food (CF); Geography, History, Language and Culture (GHLC); People and Everyday Life (PEL); Plants and Animals (PA); Science and Technology (ST); Weather and Climate (WC); and Other.

Method	Overall accuracy												
	top-1	top-3	VT	BCP	OMC	SR	CF	GHLC	PEL	PA	ST	WC	Other
Q-Only [21]	14.93	–	14.64	14.19	11.78	15.94	16.92	11.91	14.02	14.28	19.76	25.74	13.51
MLP [21]	20.67	–	21.33	15.81	17.76	24.69	21.81	11.91	17.15	21.33	19.29	29.92	19.81
BAN [44]	25.17	–	23.79	17.67	22.43	30.58	27.90	25.96	20.33	25.60	20.95	**40.16**	22.46
MUTAN [45]	26.41	–	25.36	18.95	24.02	33.23	27.73	17.59	20.09	30.44	20.48	39.38	22.46
ArticleNet (AN) [21]	5.28	–	4.48	0.93	5.09	5.11	5.69	6.24	3.13	6.95	5.00	9.92	5.33
BAN + AN [21]	25.61	–	24.45	19.88	21.59	30.79	29.12	20.57	21.54	26.42	27.14	38.29	22.16
MUTAN + AN[21]	27.84	–	25.56	23.95	26.87	**33.44**	29.94	20.71	25.05	29.70	24.76	39.84	23.62
BAN/AN oracle [21]	27.59	–	26.35	18.26	24.35	33.12	30.46	**28.51**	21.54	28.79	24.52	41.4	25.07
MUTAN/AN oracle [21]	28.47	–	27.28	19.53	25.28	35.13	**30.53**	21.56	21.68	**32.16**	24.76	41.4	24.85
GRUC (ours)	**29.87**	**32.65**	**29.84**	**25.23**	**30.61**	30.92	28.01	26.24	**29.21**	31.27	**27.85**	38.01	**26.21**

4.3.3. Experimental results on OK-VQA

We also report the quantitative performance on the challenging OK-VQA dataset in Table 3. We compare our model with three kinds of existing models, including current state-of-the-art VQA models, knowledge-based VQA models and ensemble models. The VQA models contain Q-Only [21], MLP [21], BAN [44], MUTAN [44]. The knowledge-based VQA models [21] consist of ArticleNet (AN), BAN +A N and MUTAN + AN. The ensemble models [21], i.e. BAN/AN oracle and MUTAN/AN oracle, simply take the raw ArticleNet and VQA model predictions, taking the best answer (comparing to ground truth) from either. We report the overall performance (top-1 and top-3 accuracy) as well as breakdowns for each of the knowledge categories (top-1 accuracy). We have the following two observations from the results:

First, our model consistently outperforms all the compared models on the overall performance. Even the state-of-the-art models (BAN and MUTAN) specifically designed for VQA tasks, they get inferior results compared with ours. This indicates that general VQA task like OK-VQA cannot be simply solved by a well-designed model, but requires the ability to incorporate external knowledge in an effective way. Moreover, our model outperforms knowledge-based VQA models including both single models (BAN+AN and MUTAN + AN) and ensemble models (BAN/AN oracle and MUTAN/AN oracle), which further proves the advantages of our knowledge incorporating mechanism based on both multimodal knowledge graphs and memory-enhanced recurrent reasoning network.

Second, the improvement of our model on OK-VQA is not that remarkable compared to the performance on FVQA and Visual7W-KB. We believe that this phenomenon is mostly due to the following two reasons: (1) Questions in the OK-VQA dataset are more diverse and complex, which is more challenging for machines to understand accurately. The questions in FVQA and Visual7W-KB are generated when given the images and supporting facts upon the pre-defined templates or relations. This mechanism constrains the answers in a specific knowledge base and guides the model to operate in a reverse way of the question generation process to predict the correct answers with high probability. On the contrary, questions in OK-VQA are totally free-form ones that generated by MTurk workers and thus containing more unique questions and words with less bias compared with other datasets. This increases the difficulty to understand the questions accurately. (2) OK-VQA requires a wide range of knowledge beyond a specific knowledge base. Looking at the category breakdowns in Table 3, baseline models achieve relatively high performance for SR, CF, GHLC, PA and WC categories while our model performs better for the remaining categories. Since the baseline models refer to the Wikipedia while our model refers to ConceptNet, the performance in the category breakdowns perhaps suggests that each knowledge base just provides a portion of required knowledge. To improve the overall performance on OK-VQA, we should better comprehensively consider knowledge bases that cover commonsense, visual knowledge, Wikipedia knowledge and even professional knowledge.

4.4. Ablation study

Since our model contains multiple essential components, we test a series of variations on the three benchmark datasets to verify the influence of each component. The experimental results are shown in Table 4.

4.4.1. Influence of different knowledge modalities

As demonstrated in Section 1, we believe that different modalities can provide complementary knowledge for answer inference. In this Section, we conduct ablation study to prove the indispensable role of each modality by the following variations:

- **w/o Semantic Graph (model '1')**: this model removes the semantic graph in graph construction and the follow-up reasoning process.
- **w/o Visual Graph (model '2')**: this model removes the visual graph in graph construction and the follow-up reasoning process.
- **w/o Semantic Graph & Visual Graph (model '3')**: this model simultaneously removes the visual graph and the semantic graph in graph construction and the follow-up reasoning process.

The experimental results are shown in the first block in Table 4. We observe that the top-1 and top-3 accuracy of '1' and '2' all decrease compared with the full model on the three datasets, which indicates that both semantic and visual graphs are beneficial to provide valuable evidence for answer inference. Thereinto, the visual information has greater impact than the semantic part, proving that the image content still plays essential role in KVQA tasks. When removed both semantic and visual graphs, '3' results in a significant decrease. It gives us insight that factual knowledge only is entirely insufficient to answer the question. By incorporating three of the knowledge modalities, we achieve the best performance.

4.4.2. Influence of the GRUC module

As the key component in our model, the GRUC module has two advantages over the existing reasoning models: first, it considers the structure information in the knowledge base and involves the structures in the reasoning process; second, the multi-step reasoning process via recurrent GRUC modules achieves greater reasoning

Table 4
Ablation study of key components on FVQA, Visual7W-KB and OK-VQA.

Method		FVQA		Visual7W + KB		OK-VQA	
		top-1	top-3	top-1	top-3	top-1	top-3
GRUC (full model)		**79.63**	**91.20**	**69.03**	**88.12**	**29.87**	**32.65**
1	w/o Semantic Graph	78.05	87.70	67.01	84.91	28.30	31.02
2	w/o Visual Graph	76.98	83.15	66.38	79.80	28.02	29.52
3	w/o Semantic Graph & Visual Graph	20.43	29.10	17.88	28.43	12.11	13.96
4	w/o Neighbor Aggregation (Control Unit)	78.53	89.34	68.34	85.67	28.20	30.89
5	w/o Neighbor Aggregation (Update Unit)	77.61	88.05	66.52	82.04	25.74	27.62
6	w/o GRUC Module	70.87	78.70	57.22	70.80	18.65	20.91
7	w/o Intra-Modal Knowledge Selection	74.85	80.63	67.28	85.41	26.49	27.56
8	w/o Global Assessment	79.10	90.54	68.43	87.69	29.80	32.11

ability compared with the state-of-the-art OB model [6] merely applying feature concatenation. We justify these two advantages by the following variations:

- **w/o Neighbor Aggregation (Control Unit) (model '4')**: this model removes the neighborhood information, i.e. $\boldsymbol{c}_i^F$ in Eq. (5), in initializing the control unit.
- **w/o Neighbor Aggregation (Update Unit) (model '5')**: this model removes the neighborhood information, i.e. $\boldsymbol{c}_j^{nei}$ in Eq. (11), in the update unit.
- **w/o GRUC network (model '6')**: this model replaces the GRUC network and multimodal feature fusion introduced in Section 3.3 by direct concatenation, *i.e.* concatenating the mean pooling of all the semantic/visual node features with each entity feature.

The experimental results are shown in the second block in Table 4. The performance on three datasets decreases slightly when remove the neighborhood information from either the control unit or the update unit. It indicates that preserving the structural information when incorporating the knowledge brings richer semantics for answer prediction. The performance decreases more than 10% when replacing the GRUC network by simple concatenation, which proves the advantages of the proposed recurrent reasoning process in gathering complementary evidence from different modalities.

4.4.3. Influence of the intra-modal knowledge selection

The first stage in our model is to select knowledge from each modalities under the guidance of the question independently. Since most questions are referring to a small portion of knowledge, this stage aims to choose relevant knowledge and avoid unexpected noise for improving the performance. We prove this motivation by the variation below:

1. - **w/o Intra-Modal Knowledge Selection (model '7')**: this model removes the intro-modal knowledge selection process introduced in Section 3.2.

The experimental results are shown in the third block in Table 4. We observe that all the metrics decrease remarkably on all the three datasets. It indicates that 'intuitive' knowledge filtering before 'clever' knowledge reasoning is effective to bring extra improvement.

4.4.4. Influence of the global assessment

The last stage in our model is to globally assess all the concepts via GNN and choose the optimal one as the answer. To prove the influence of this process, we further conduct the following ablation study:

- **w/o Global Assessment (model '8')**: this model removes the GNN operated on all the concepts in Section 3.4 and feeds the embedding of each concept in $\{\widetilde{\boldsymbol{v}}_i^F\}_{i=1}^{\mathcal{N}^F}$ directly to the binary classifier.

The experimental results are shown in the last block in Table 4. The performance consistently decreases on all the datasets. However, the decrease is relatively smaller compared with other models. We think that the GNN model used in this process is relatively simple and more effective global assessment approach perhaps can bring more improvement.

4.5. Interpretability

Our model is interpretable by visualizing the attention weights and gate values in the cross-modal heterogeneous graph reasoning process. From case study in Figs. 6 and 7, we conclude with the following three insights:

GRUC is capable to reveal the knowledge selection mode from different modalities. In Fig. 6, the first four examples indicate that GRUC captures the most relevant visual, semantic and factual evidence (according to intra-modal attention weights) as well as complementary information across modalities (according to attention weights in the Read Unit of the GRUC module). The *ratio of total gate values* reveals the amount of information derived from each graph. We denote the sum of all the gate values for the visual dimensions, fact dimensions and semantic dimensions in Eq. (13) as G_v, G_f and G_s, respectively. We denote the sum of all the gate values for all the dimensions in Eq. (13) as G. The *ratio of total gate values* for the visual graph, fact graph and semantic graph is defined as $\frac{G_v}{G}$, $\frac{G_f}{G}$ and $\frac{G_s}{G}$, respectively. In most cases, factual knowledge provides predominant clues compared with other modalities from the observation of the ratio of gate values. It is because that KVQA tasks rely on external knowledge to a great extent. Furthermore, more evidence comes from the semantic modality when the question involves complex relationships. For instance, the question in the first case asking about single object 'devise' requires more visual information while the question in the second case involving the relationship between 'hand' and 'while round thing' needs more semantic clues.

GRUC has advantages over the state-of-the-art model. In Fig. 6, the fifth example compares the predicted answer of Out of the Box (OB) [6] with GRUC. GRUC collects relevant visual and semantic evidence to make each entity discriminative enough for predicting the correct answer while OB failing to distinguish representations of 'laptop' and 'keyboard' without feature selection.

GRUC fails mostly in three conditions: highly relevant answers, inadequate visual evidence and limited external knowledge. Fig. 6 shows four failure cases. (1) Some cases fail when the predicted answer is quite relevant to the ground truth answer. In the first case, it's reasonable that both 'wedding' and 'party' may have cakes. There is no further evidence from the image to decide which situation is more accurate. We can explain the second case by the similar reason. (2) Some cases fail when there is no enough evidence to infer the correct answer, such as the third sample in Fig. 6. (3) Some other failure cases are due to the lack of required

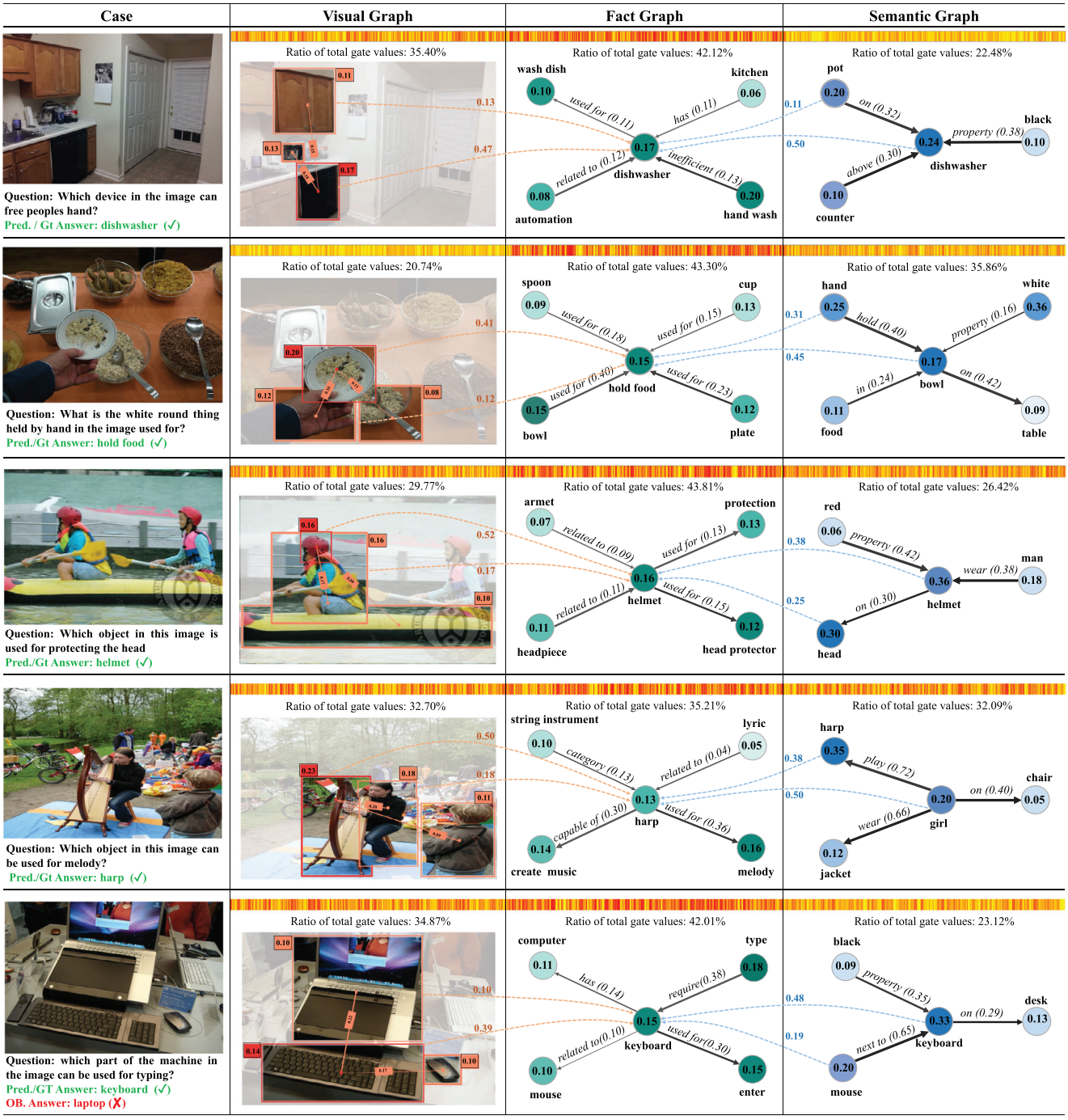

Fig. 6. Visualization for our model. Visual graph highlights the most relevant subject (red box) according to attention weights of each object (α^V in Eq. (1)) and the objects (orange boxes) with top-2 attended relationships (β^V in Eq. (2)). Fact graph shows the predicted entity (center node) and its top-4 attended neighbors (α^F in Eq. (1)). Semantic graph shows the most relevant concept (center node) and its up to top-4 attended neighbors (α^S in Eq. (1)). Each edge is marked with attention value ($\beta^{F|S}$ in Eq. (2)). Dash lines represent memory read attention weights at reasoning step T ($\gamma^{(T)}$ in Eq. (9)). The thermogram on the top visualizes the gate values ($\boldsymbol{gate}_i$ in Eq. (13)) of visual embedding (left), entity embedding (middle) and semantic embedding (right). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

knowledge in the provided knowledge base. In the last case in Fig. 6, there is no fact about 'samuel fox' in ConceptNet. Therefore, comprehensively considering multiple knowledge bases to cover a wider range of knowledge is important to improve the KVQA ability.

4.6. Parameter analysis

We evaluate the influence of different number of dense captions in Table 5. The results show that 10 captions achieves the best performance on both top-1 and top-3 accuracy. We further evaluate the influence of different number of reasoning steps T in the GRUC network. Fig. 8 shows the top-1 and top-3 accuracy

Table 5
Overall accuracy with different number of dense captions.

#top-k dense captions	5	10	20
top-1 accuracy	70.20	**73.06**	65.40
top-3 accuracy	82.65	**85.94**	75.98

Fig. 7. Visualization of four failure cases.

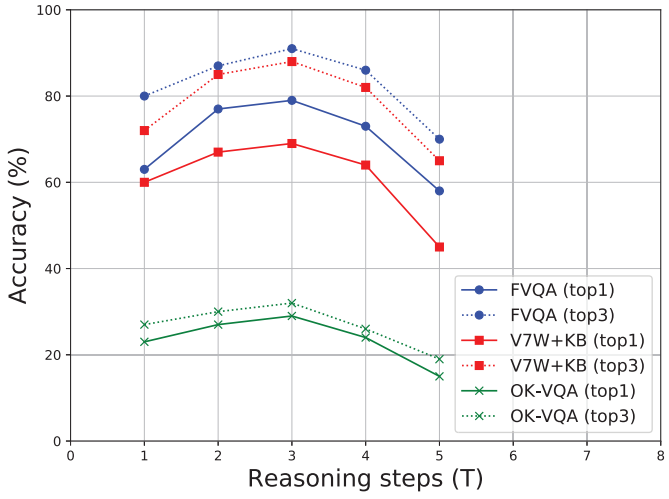

Fig. 8. Overall accuracy with different number of reasoning steps in the GRUC network. The blue solid and dotted lines respectively denote the top-1 and top-3 accuracy on FVQA. The red solid and dotted lines respectively denote the top-1 and top-3 accuracy on Visual7W+KB. The green solid and dotted lines respectively denote the top-1 and top-3 accuracy on Ok-VQA. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

on FVQA, Visual7W + KB and OK-VQA when setting the number of reasoning steps in the range of 1 to 5. We find that 3 reasoning steps achieve the best performance on the all datasets. If the number of reasoning steps less than 3, the GRUC network cannot extract adequate knowledge from each memory to support the global assessment. In contrast, too many reasoning steps may lead to over-smoothing, leading to the features of nodes converging to the similar values. Therefore, we use this setting in our full model and the ablation models.

5. Conclusion

In this paper, we propose a graph-based recurrent reasoning network GRUC for visual question answering requiring external knowledge, which focuses on cross-modal knowledge reasoning upon graph-structured multimodal knowledge representations. We novelly depict multimodal knowledge sources by multiple knowledge graphs from the visual, semantic and factual views. The representations of different modalities are unified in graph domain, thus benefiting for relational reasoning across modalities. Meanwhile, introducing the semantic graph for high-level abstraction brings remarkable improvement in KVQA, which has been less studied in previous work. On top of these representations, we propose a new recurrent reasoning model and each reasoning step is performed by a Graph-based Read, Update, and Control (GRUC) module that conducts parallel reasoning over both visual and semantic information. GRUC is a parallel reasoning module that applies modality-oriented controllers for reasoning over different modalities in a parallel mode, which can be easily extended to involve more modalities. Our model consistently outperforms the state-of-the-art approaches remarkably on FVQA, Visual7W-KB and OK-VQA datasets. Furthermore, the model has good interpretability of revealing the knowledge selection mode from different modalities by comprehensive visualization. However, our model has inferior performance when open-domain knowledge is required. How to adaptively incorporate diverse knowledge bases that covering commonsense, Wikipedia knowledge and even professional knowledge for KVQA tasks will be our future work.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Acknowledgment

This work was supported by the National Key Research and Development Program of China under grant no. 2016YFB0801003.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, D. Parikh, VQA: visual question answering, in: ICCV, 2015, pp. 2425–2433.
[2] Z. Fang, J. Liu, Y. Li, Y. Qiao, H. Lu, Improving visual question answering using dropout and enhanced question encoder, Pattern Recognit. 90 (2019) 404–414.
[3] J. Yu, W. Zhang, Y. Lu, Z. Qin, Y. Hu, J. Tan, Q. Wu, Reasoning on the relation: enhancing visual representation for visual question answering and cross-modal retrieval, TMM (2020), doi:10.1109/TMM.2020.2972830.
[4] P. Wang, Q. Wu, C. Shen, A. Dick, A. van den Hengel, FVQA: fact-based visual question answering, PAMI 40 (10) (2018) 2413–2427.
[5] P. Wang, Q. Wu, C. Shen, A.R. Dick, A. van den Hengel, Explicit knowledge-based reasoning for visual question answering, in: IJCAI, 2017, pp. 1290–1296.
[6] M. Narasimhan, S. Lazebnik, A. Schwing, Out of the box: reasoning with graph convolution nets for factual visual question answering, in: NeurIPS, 2018, pp. 2654–2665.
[7] J.L. McClelland, F. Hill, M. Rudolph, J. Baldridge, H. Schütz, Extending machine language models towards human-level language understanding, arXiv preprint arXiv:1912.05877 (2019).
[8] M. Ren, R. Kiros, R. Zemel, Exploring models and data for image question answering, in: NeurIPS, 2015, pp. 2953–2961.
[9] Z. Yu, J. Yu, J. Fan, D. Tao, Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, in: ICCV, 2017, pp. 1821–1830.
[10] H. Ben-Younes, R. Cadene, N. Thome, M. Cord, Block: bilinear superdiagonal fusion for visual question answering and visual relationship detection, in: AAAI, 2019, pp. 8102–8109.
[11] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in: NeurIPS, 2016, pp. 289–297.
[12] W. Wang, Y. Huang, L. Wang, Long video question answering: a matching-guided attention model, Pattern Recognit. 102 (2020).
[13] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: CVPR, 2018, pp. 6319–6328.
[14] A. Santoro, D. Raposo, D.G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, T. Lillicrap, A simple neural network module for relational reasoning, in: NeurIPS, 2017, pp. 4967–4976.

[15] W. Norcliffe-Brown, S. Vafeias, S. Parisot, Learning conditioned graph structures for interpretable visual question answering, in: NeurIPS, 2018, pp. 8334–8343.
[16] C. Wu, J. Liu, X. Wang, X. Dong, Chain of reasoning for visual question answering, in: Advances in Neural Information Processing Systems, 2018, pp. 275–285.
[17] D. Hudson, C.D. Manning, Learning by abstraction: the neural state machine, in: NeurIPS, 2019, pp. 5901–5914.
[18] X. Jiang, J. Yu, Z. Qin, Y. Zhuang, X. Zhang, Y. Hu, Q. Wu, Dualvd: an adaptive dual encoding model for deep visual understanding in visual dialogue, AAAI, 2020.
[19] M. Narasimhan, A.G. Schwing, Straight to the facts: learning knowledge base retrieval for factual visual question answering, in: ECCV, 2018, pp. 451–468.
[20] G. Li, H. Su, W. Zhu, Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks, arXiv preprint arXiv:1712.00733 (2017).
[21] K. Marino, M. Rastegari, A. Farhadi, R. Mottaghi, OK-VQA: a visual question answering benchmark requiring external knowledge, in: CVPR, 2019, pp. 3195–3204.
[22] F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Trans. Neural Netw. 20 (1) (2009) 61–80.
[23] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, ICLR, 2017.
[24] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, ICLR, 2017.
[25] R. Liao, Z. Zhao, R. Urtasun, R. Zemel, Lanczosnet: multi-scale deep graph convolutional networks, ICLR, 2019.
[26] Y. Li, D. Tarlow, M. Brockschmidt, R.S. Zemel, Gated graph sequence neural networks, ICLR, 2016.
[27] M. Schlichtkrull, T.N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, M. Welling, Modeling relational data with graph convolutional networks, in: ESWC, 2018, pp. 593–607.
[28] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, P.S. Yu, Heterogeneous graph attention network, in: WWW, 2019, pp. 2022–2032.
[29] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks 39 (6) (2017) 1137–1149.
[30] J. Johnson, A. Karpathy, L. Fei-Fei, DenseCap: fully convolutional localization networks for dense captioning, in: CVPR, 2016, pp. 4565–4574.
[31] P. Anderson, B. Fernando, M. Johnson, S. Gould, Spice: semantic propositional image caption evaluation, in: ECCV, 2016, pp. 382–398.
[32] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Neural message passing for quantum chemistry, in: ICML, 2017, pp. 1263–1272.
[33] Y. Li, D. Tarlow, M. Brockschmidt, R.S. Zemel, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: EMNLP, 2014, pp. 1724–1734.
[34] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, in: NeurIPS, 2017, pp. 1024–1034.
[35] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, DBPedia: a nucleus for a web of open data, in: The Semantic Web, Springer, 2007, pp. 722–735.
[36] R. Speer, J. Chin, C. Havasi, ConceptNet 5.5: an open multilingual graph of general knowledge, in: AAAI, 2017, pp. 4444–4451.
[37] N. Tandon, G. De Melo, F. Suchanek, G. Weikum, WebChild: harvesting and organizing commonsense knowledge from the web, in: WSDM, 2014, pp. 523–532.
[38] Y. Zhu, O. Groth, M. Bernstein, L. Fei-Fei, Visual7w: grounded question answering in images, in: CVPR, 2016, pp. 4995–5004.
[39] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, et al., Visual genome: connecting language and vision using crowdsourced dense image annotations, IJCV 2017 (123) (2017) 32–73.
[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: ECCV, Springer, 2014, pp. 740–755.
[41] J. Pennington, R. Socher, C.D. Manning, GloVe: global vectors for word representation, in: EMNLP, 2014, pp. 1532–1543.
[42] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.
[43] H. Li, P. Wang, C. Shen, A. van den Hengel, Visual question answering as reading comprehension, in: CVPR, 2019, pp. 6319–6328.
[44] J.-H. Kim, J. Jun, B.-T. Zhang, Bilinear attention networks, in: NeurIPS, 2018, pp. 1564–1574.
[45] H. Ben-Younes, R. Cadene, M. Cord, N. Thome, MUTAN: multimodal tucker fusion for visual question answering, in: ICCV, 2017, pp. 2612–2620.

Jing Yu is an Assistant Professor in the Institute of Information Engineering, Chinese Academy of Sciences. She received her B.S. degree in Automation Science from Minzu University, China, in 2011, and got her M.S. degree in Pattern Recognition from Beihang University, China in 2014. She received her Ph.D. degree in the University of Chinese Academy of Sciences, China, in 2019. She works on Vision and Language problems, including visual question answering, visual dialogue, cross-modal information retrieval, etc.

Zihao Zhu received his B.S. degree in Computer Science and Technology from China University of Mining and Technology in 2018. Now he is a graduate student in Institute of Information Engineering, Chinese Academy of Sciences, China and School of Cyber Security, University of Chinese Academy of Sciences, China. His research interests include multimodal machine learning, visual question answering.

Yujing Wang is a senior researcher in Microsoft Research Asia. She received her B.S. and M.S. degrees from Peking University, China, in 2010 and 2013, respectively. Her current research interests include Automated Machine Learning, Natural Language Understanding, and Graph Mining.

Weifeng Zhang received his B.S degree in Electronic Information Engineering from Beijing University of Technology in 2009, got his M.S degree in Pattern Recognition from Beihang University in 2012, and got his Ph.D. degree in Computer Science from Hangzhou Dianzi University in 2019. Now, he is an associate professor in Jiaxing University. His research interests include machine learning, multimedia modeling.

Yue Hu is a Professor in the Institute of Information Engineering, Chinese Academy of Sciences. She received her Ph.D. in computer science from University of science and technology, Beijing in 2000. She is currently a researcher in Institute of Information Engineering, Chinese Academy of Sciences. Her research interest are in the area of natural language processing and social network analysis.

JianLong Tan received his Ph.D. degree in Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2003. He received his B.S. and M.S. degree in Xiangtan University, China, in 1997 and 2000, respectively. He is currently a professor in Institute of Information Engineering, Chinese Academy of Sciences. His research interests are in the area of multimedia analysis and hardware algorithm design.