Neurocomputing 488 (2022) 54–65

Contents lists available at ScienceDirect

Neurocomputing

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

You should know more: Learning external knowledge for visual dialog

Lei Zhao [a,b], Haonan Zhang [a], Xiangpeng Li [a], Sen Yang [a], Yuanfeng Song [a,*]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China
[b] Sichuan Artificial Intelligence Research Institute, Yibin, China

ARTICLE INFO

Article history:
Received 8 June 2021
Revised 19 October 2021
Accepted 31 October 2021
Available online 12 February 2022

Keywords:
External knowledge
Graph convolutional network
Structured knowledge graph
Attention mechanism
Visual dialog

ABSTRACT

Visual dialog is a task that two agents complete a multi-round conversation based on an image, a caption, and dialog histories. Despite the recent progress, existing methods still undergo degradation on the condition of complex scenarios. Handling these scenarios depends on logical reasoning that requires commonsense priors. In this paper, we propose a novel visual dialog pipeline named Structured Knowledge-Aware Network (SKANet), consisting of an Image Knowledge-Aware Module and a Caption Knowledge-Aware Module. Specifically, the Image and Caption Knowledge-Aware Modules construct commonsense knowledge graphs from ConceptNet. We apply SKANet to two sub-tasks: the conventional visual dialog and a goal-oriented visual dialog named 'image guessing'. For the conventional visual dialog, the SKANet is combined with an additional Multi-Modality Fusion Module, which is designed to explore the visual content and the textual context about the dialog history. For the goal-oriented visual dialog, we directly apply the Image and Caption Knowledge-Aware Modules to two agents, respectively. Experimental results on VisDial v0.9 and VisDial v1.0 datasets show that our proposed method effectively outperforms comparative methods on both sub-tasks.

1. Introduction

Currently, bridging the gap between vision and language, such as visual captioning [1–4], cross-modal retrieval [5–7], visual question answering [8–11], and visual dialog [12–15], has attracted massive attention in the computer vision community. Visual dialog can be regarded as an extension of VQA. It accomplishes a multi-round conversation depending on an image, a caption, and dialog histories. As the Visual Turing Test, visual dialog can be applied to various scenarios, including visual aids and robotics.

Since Das et al. [16] collected the visual dialog dataset (VisDial) and proposed the baseline that simply fuses the multi-modality data, many related methods [17,13,18,19] were proposed to answer the stationary questions. Its essence is to complete visual grounding [20–23] according to the text information. Besides this traditional visual dialog, the goal-oriented visual dialog [24–26] were also put forward. It simultaneously generates questions and needs to achieve a specific sub-task by the generated dialogs, such as image guessing. During the development of these sub-tasks, most of them focus on the attention mechanism [27] and the graph model [28] to explore the internal features about the image and the dialog history. However, it is insufficient to handle complex conversations only with the internal knowledge. Human beings need much common sense to construct auxiliary reasoning when dealing with daily conversation. As for the visual dialog task, it also needs logical reasoning with commonsense knowledge to construct semantic relevance among objects. As shown in Fig. 1, the external knowledge provides potential related common sense to answer the question.

In this paper, we propose a novel Structured Knowledge-Aware Network (SKANet), which contains an Image Knowledge-Aware Module and a Caption Knowledge-Aware Module. The Image and Caption Knowledge-Aware Modules explore the commonsense knowledge graph from the image and the caption, respectively. The process of commonsense knowledge graph encoding contains three steps: 1) concept recognition. The concept set of the image is extracted by Faster R-CNN [29] trained on Visual Genome [30], and the concept set of the caption is recognized by matching tokens from ConceptNet [31]; 2) graph construction. The construction for both the caption and the image is implemented by sub-graph matching and path pruning. The sub-graph of ConceptNet covers all the concept pairs in the caption and the image. And it provides the relevant knowledge for question answering. After that, the low-quality paths are pruned by a knowledge graph embedding technique; and 3) graph feature extraction. The pruned graph is encoded by a graph convolutional network (GCN) [32] to extract structured features.

[*] Corresponding author.
E-mail address: songyf@uestc.edu.cn (Y. Song).

https://doi.org/10.1016/j.neucom.2021.10.121
0925-2312/© 2022 Elsevier B.V. All rights reserved.

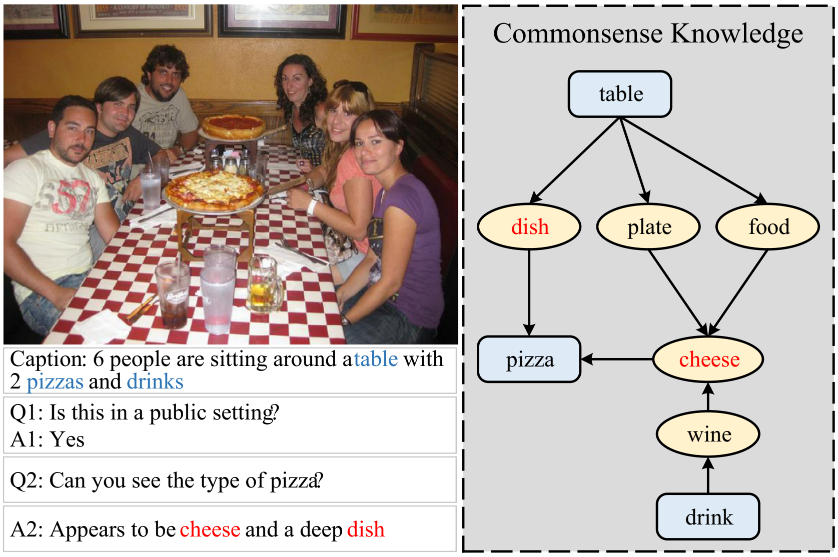

Caption: 6 people are sitting around a table with 2 pizzas and drinks
Q1: Is this in a public setting?
A1: Yes
Q2: Can you see the type of pizza?
A2: Appears to be cheese and a deep dish
Commonsense Knowledge
table
dish
plate
food
pizza
cheese
wine
drink

Fig. 1. An example that shows the effect of commonsense knowledge extracted from ConceptNet for visual dialog. The words in blue, including "table","pizza", "drink" are the concepts grounded from the caption. The words in the light yellow ellipse, like "dish", "plate", "food", represent the related concepts. The words in red, like "cheese", are the valuable concepts for the question. When answering the question about "the type of pizza", the commonsense knowledge can serve as an effective complement to answer "cheese" or "dish".

To validate the effectiveness of SKANet, we apply it to the conventional visual dialog and the goal-oriented visual dialog named 'image guessing'. For the conventional visual dialog, we design an additional Multi-Modality Fusion Module to explore the textual features on the dialog history and visual features on the image with the attention mechanism. Then the structured graph features from the Knowledge-Aware Modules are separately queried by the embedded features of the question. Finally, all the features are fused to generate an answer. Unlike the conventional visual dialog that only needs answer the questions, the cooperative 'image guessing' is played by Question-BOT (Q-BOT) and Answer-BOT (A-BOT) on the condition of information asymmetry. The asymmetry means that A-BOT can see the input image, but Q-BOT only knows its caption. Meanwhile, Q-BOT generates natural questions, and A-BOT is in charge of answering them. At the end of each round of dialog, Q-BOT predicts a representation of the unseen image and guesses the corresponding image from the candidate list by the distance similarity. We apply those two modules of SKANet to Q-BOT and A-BOT, separately. The Image Knowledge-Aware Module helps A-BOT to solve complex scenes like the conventional visual dialog task. Furthermore, the Caption Knowledge-Aware Module helps Q-BOT to generate refined question and image representation. Finally, curriculum learning combined with supervised pretraining and deep reinforcement learning is used to learn the policies of these two bots.

The contributions of this paper are concluded as follows:

- To supplement the external common sense missed in visual dialog, we propose a Structured Knowledge-Aware Network (SKANet) by incorporating the commonsense knowledge derived from ConceptNet.
- We construct structured knowledge on the image and the caption to capture the semantic relevance among objects. To extract the relational context among entities of the graph, GCN is applied to encode the knowledge graph.
- To demonstrate the effectiveness of SKANet, we apply it to two sub-tasks: the conventional visual dialog and the goal-oriented visual dialog named 'image guessing'. Experiments conducted on VisDial v0.9 and VisDial v1.0 show that our proposed method outperforms comparative approaches, and an ablation study further verifies the effectiveness of our model.

The rest of this paper is organized as follows. The related work is introduced in Section 2. The detailed methods are described in Section 3. The extensive Experiments and their results are presented in Section 4. Finally, a conclusion about this work is drawn in Section 5.

2. Related work

In this section, the related works about this work are presented as follows: visual question answering, visual dialog, and external knowledge base.

2.1. Visual question answering

Visual question answering (VQA) [33] is a multi-modal task that aims to answer an accurate answer on the basis of an image and a visually relevant question. Different from the textual question answering, visual and textual information both need to be understood and reasoned in the VQA. It is the most challenging difficulty that how to integrate the multi-modal data effectively. The joint embedding methods [34,35] directly embed the image and the question into the same semantic space. However, the effective regional features can not be extracted. Then many attention-based methods [1,36–38] are proposed. The attention mechanism makes the models focus on the question-related regions. Specifically, Nguyen et al. [39] propose a symmetric module between visual and textual representation. The question and visual region interact with each other in this module. Yu et al. [38] apply self-attention and guided-attention units to compose a modular co-attention layer, which can further deepen the interaction.

2.2. Visual dialog

Visual dialog aims to accomplish ten rounds of conversation depending on an image, a caption, and the corresponding dialog history. We divide this task into two sub-tasks.

The conventional visual dialog focuses on the question answering. Most of methods solve the task by attention mechanism or graph construction. Kang et al. [19] use two attention modules named REFER and FIND to solve visual reference resolution. The REFER applies a multi-head attention module to excavate valuable history representation. The FIND is responsible for the further visual grounding. Niu et al. [40] propose Recursive Visual Attention to infer co-reference in dialog. The retrospection of history would continue until the visual co-reference is determined with high confidence. Attention mechanism is a powerful vehicle for capturing the significant features, but it is incapable in face of constructing complex potential relationship among the inputs. Graph model can somewhat handle this case. Schwartz et al. [13] apply a factor graph based attention to integrating all of the various features. The factors in the graph model the interactions among the utilities. Guo et al. [41] propose a context-aware graph network to further enhance the integration capability. Moreover, the graph model is updated by an adaptive top-K message passing mechanism. These graph-based methods are usually complicated. Meanwhile, some methods [42,43] that also put forward questions are then proposed. Jain et al. [42] propose a symmetric discriminative baseline to predict answers and questions. Conditional VAE is then applied to answer questions in [43].

The goal-oriented visual dialog consists of ‘GuessWhat’ [24,44,25] and ‘image guessing’ [26,45,46]. They are both a two-agent guessing game. In ‘GuessWhat’, Q-BOT which was responsible for generating questions, locates the intended object in the input image. In ‘image guessing’, Q-BOT aims to select an unseen image from a gallery of images. After each round of dialog, the result of the guess is regarded as a reward. In this work, we apply SKANet to the traditional visual dialog and the ‘image guessing’ task.

2.3. External knowledge base

There are many external knowledge bases that have been proposed in the past years, such as Dbpedia [47], ConceptNet [31], ATOMIC [48], GenericsKB [49]. Dbpedia extracts structured and rich knowledge from Wikipedia. ConceptNet is produced from the Open Mind Common Sense(OMCS) corpus based on the WordNet. Its knowledge is more natural compared with other external knowledge bases. ConceptNet 5.5 consists of about 8 million nodes which have 21 million edges. ATOMIC focuses on the inferential knowledge whose relationship is about ‘if and then’. It includes 877 k pieces of textual reasoning knowledge. GenericsKB is the first large resource that contains naturally existing generic sentences instead of the crowdsourced or extracted triplets. Its captions are of high quality.

The external knowledge have been wildly applied to the field that bridges the computer vision and NLP, such as VQA. Shah et al. [50] propose the first knowledge-aware dataset KVQA for VQA. However, the external knowledge base have been rarely adopted in the visual dialog task. In this paper, we employ ConceptNet 5.5 as the external knowledge provider. The image and the corresponding caption separately generate the concept sets, which are applied to extract the related commonsense knowledge. This external knowledge can help our model handle many complex scenarios.

3. Methodology

Visual dialog aims to complete a multi-round conversation between two agents. To deal with the complex scenarios, we propose a novel Structured Knowledge-Aware Network (SKANet) to improve the reasoning ability. SKANet is applied to two sub-tasks: the conventional visual dialog and the goal-oriented visual dialog named ‘image guessing’. In this section, we introduce the particular application of SKANet on these sub-tasks.

3.1. Task 1: conventional visual dialog

Given an image I with its corresponding caption C and a dialog history $H_t = [(q_1, a_1), (q_2, a_2), \ldots, (q_{t-1}, a_{t-1})]$, the conventional visual dialog task aims to infer an answer for the current question q_t by sorting a list of 100 candidate answers. Our proposed method (SKANet) for this sub-task is illustrated in Fig. 2. It mainly consists of three modules. Besides the direct propagation of textual features on the dialog history and visual features on the image (the Multi-Modality Fusion Module), the grounded concepts of the image and the caption construct sub-graphs from ConceptNet, respectively (the Image Knowledge-Aware Module and the Caption Knowledge-Aware Module).For the Multi-Modality Fusion Module, the attention mechanism is utilized to ground the relevant visual object depending on the question and the dialog history. For the Image Knowledge-Aware Module, we want to apply the related external knowledge about the image to improve the reasoning ability of the model. Since the caption of the input image may contain more information, the external knowledge about the caption is used to further enhance the performance. The following subsections describe the details of SKANet for this sub-task.

3.1.1. Preliminaries

In this section, we introduce the preparation for our method. It describes the process of data. There are two types of input, including textual data and visual data. The textual data consists of the current question q_t, the caption C, the historical dialog H, and the candidate answer options A. Generally, the caption C also acts as a part of history information. The visual data refers to the input image I. In terms of the input object of the data, the history dialog H and the image I are input into the Multi-Modality Fusion Module. The caption C and the image I are input into Knowledge-Aware Module. The current question q_t is input into the Multi-Modality Fusion Module and the Knowledge-Aware Module to generate question-related representations.

The current question q_t is first embedded by the pre-trained GloVe embeddings [51]. Then the embedded vector $W^q = (w_1, \ldots, w_n) \in \mathbb{R}^{d_W \times n_q}$ is fed into a BiLSTM to encode a sequential representation $s^q \in \mathbb{R}^{d_S \times 1}$ that is used for the Question-Aware Attention. n_q denotes the number of tokens in the question q_t.

The history dialog H_t **and the candidate answer options** are also encoded by different BiLSTMs. Specially, the encoded sequential representation of the history dialog is denoted as $S_t^h = (p_0, p_1, \cdots, p_{t-1}) \in \mathbb{R}^{d_S \times t}$, where p_0 is the feature of caption C, p_r is the feature of the r-th ($r \in \{1, 2, \ldots, t-1\}$) round question–answer pair.

The image I is fed into Faster R-CNN that is pre-trained on Visual Genome to generate object-level features and labels of the objects. The fixed 36 proposal features are encoded into a sequence representation $V = (v_1, v_2, \ldots, v_{n_p})$, which is input into the Multi-Modality Fusion Module. n_p denotes the index of the proposal. The labels of the objects are integrated into a set, such as {man, skirt, woman, table, pizza, water}, and then we feed it into the Image Knowledge-Aware Module to construct the sub-graph.

The caption C for the Caption Knowledge-Aware Module is processed by concept grounding, which is a process of token matching from ConceptNet.

The ConceptNet is the only supplier for the external knowledge. It uses triples (*start, relation, end*) to connect concepts with the relationship between each other. For example, “*a bird has wings*” is represented as (*bird, HasA, wings*). “Bird” is the start node, “wings” is the end node, and “HasA” is the edge that represents

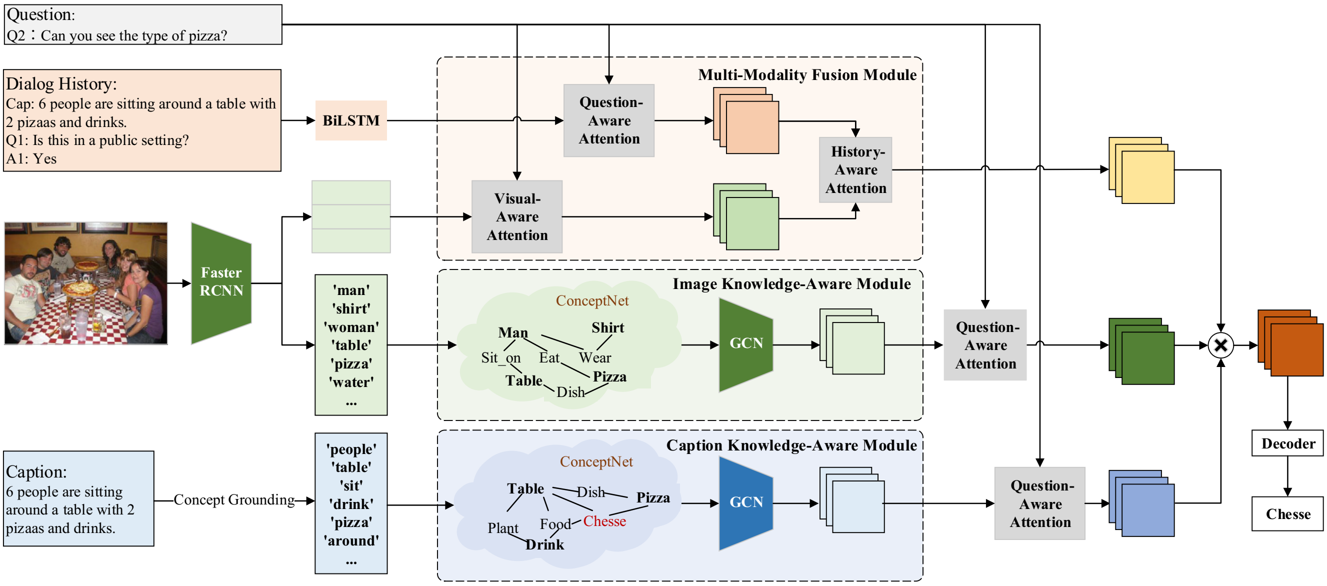


Fig. 2. Overview of our proposed framework SKANet for the conventional visual dialog task. We adopt the traditional encoder-decoder model. The encoder mainly consists of the Multi-Modality Fusion Module, the Image Knowledge-Aware Module, and the Caption Knowledge-Aware Module. The output features of the Knowledge-Aware Module are separately queried by the embedding feature of the question. Then the attentive features are fused and feed into the decoder to produce the final answer.

the relationship between the start node and the end node. There are nearly 1.5 million nodes about English vocabulary in ConceptNet. It can provide lots of common sense to help the answer agent generate precise answers like humans. In the visual dialog task, we only need to justify whether there is a relationship between two nodes. All of the edge types are ignored for simplicity.

The matching mechanism is to match n-grams in the caption C with the concept set. Following [52], we apply soft matching with lemmatization and filtering of stop words to improve the naive method. After the concept grounding of the caption "*A woman sits on a bench holding a guitar in her lap*", the concept set {woman, sit, sit_on, bench, guitar, lap} is produced.

3.1.2. Multi-modality fusion module

The Multi-Modality Fusion Module mainly consists of three Attention Modules: Question-Aware Attention, Visual-Aware Attention, and History-Aware Attention. The question guides the history dialog by sentence-level sequential representation to generate the features that are relevant to the question. The visual features query the token-level question representation to get the visually related topics. Then the attentive token-level question features are concatenated with the visual features. In order to alleviate the co-reference, the concatenated features are then queried by the question-related history features.

The sequence feature of the current question s^q queries the dialog history S_t^h as follows:

$$\begin{aligned} \alpha^h &= \mathrm{softmax}\left(\mathrm{FC}\left(\mathrm{f}_q^h(s^q) \circ \mathrm{f}_p^h\left(S_t^h\right)\right)\right), \\ \tilde{s}^h &= \sum_{r=0}^{t-1} \alpha_r^h p_r, \end{aligned} \tag{1}$$

where $\circ$ denotes element-wise multiplication, $\mathrm{f}_q^h(\cdot)$ and $\mathrm{f}_p^h(\cdot)$ are non-linear layers, $\mathrm{FC}(\cdot)$ is a linear layer. Then a gate function with sigmoid activation is used to integrate the attentive history features as follows:

$$\begin{aligned} g^h &= \sigma\left(\mathrm{FC}\left(\left[s^q, \tilde{s}^h\right]\right)\right), \\ e^h &= g^h \circ \left[s^q, \tilde{s}^h\right], \end{aligned} \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function. e^h is an input of History-Aware Attention.

Meanwhile, we apply the Visual-Aware Attention to fetch the more effective question features. The visual representation V queries the token-level sequential representation of the question W^q. Specifically, a dot-product attention and an MLP (multi-layer perceptron) are used as follows:

$$\begin{aligned} r_{ik}^v &= \mathrm{f}_q^v(w_i)^\top \mathrm{f}_p^v(v_k), \\ \alpha_{ik}^v &= \exp\left(r_{ik}^v\right) / \sum_{j=1}^{n_q} \exp\left(r_{jk}^v\right), \\ w_k^v &= \mathrm{MLP}\left(\left[v_k, \sum_{i=1}^{n_q} \alpha_{ik}^v w_i\right]\right), \end{aligned} \tag{3}$$

where $\mathrm{f}_q^v(\cdot)$ and $\mathrm{f}_p^v(\cdot)$ are non-linear layers.

Then the question-related history feature e^h queries the visually grounded question features w_k^v with soft attention and gated function, which is called History-Aware Attention. The details are as follows:

$$\begin{aligned} r_k^h &= \mathrm{f}_q^u\left(w_k^v\right)^\top \mathrm{f}_p^u\left(e^h\right), \\ \alpha_k^h &= \exp\left(r_k^h\right) / \sum_{j=1}^{n_p} \exp\left(r_j^h\right), \\ F^{vh} &= \sum_{k=1}^{n_p} \alpha_k^h w_k^v, \end{aligned} \tag{4}$$

where $\mathrm{f}_q^u(\cdot)$ and $\mathrm{f}_p^u(\cdot)$ are also non-linear layers.

Finally, the output features F^{vh} that effectively integrate multi-modal representations are produced.

3.1.3. Knowledge-aware module

The Knowledge-Aware Module is composed of the Image Knowledge-Aware Module and the Caption Knowledge-Aware Module. Compared with the input image, its caption refines the image information. It also contains more latent information, including subjective statements and adjectives. The external commonsense knowledge is provided by ConceptNet that is a knowledge graph *G*. It generates rich semantic relevance that can assist the question answering. There are two steps in both of these modules: 1) graph construction; and 2) GCN encoding.

Graph construction. The graph nodes are the concepts extracted from the image and the caption, which are described in the preliminary section. The graph construction has threefold: concept pairs extracting, path finding, and path pruning.

Concept pairs extracting aims to generate a set of concept pairs. As we want to construct a directed graph, the m-th concept c_m and the n-th concept c_n are grouped into two concept pairs as (c_m, c_n) and (c_n, c_m).

Path finding is to check whether there is a path whose length is shorter than four between c_m and c_n. In this way, all of the satisfactory paths are retained.

Path pruning is to discard the redundant and low-quality data. The paths are decomposed as triplets scored by TransE [53]. We select the paths whose score exceed the threshold that is set as 1.5. Finally, the pruned paths are constructed as the corresponding graph G^v and G^c for the image and the caption, separately.

GCN Encoding. The constructed graph is encoded by GCN [32]. It extracts the relational context among entities of the graph. The concepts in the image graph G^v and the caption graph G^c are separately embedded as E^v and E^c by the TransE mentioned above. Then the visual knowledge features F_v^g and the caption knowledge features F_c^g are generated as follows:

$$F_v^g = \mathrm{GCN}(G^v, E^v), \\ F_c^g = \mathrm{GCN}(G^c, E^c). \tag{5}$$

The visual knowledge features F_v^g and the caption knowledge features F_c^g are then queried by the question q_t to focus on the question-related features. Their attention module is the same as the Question-Aware Attention in Multi-Modality Fusion Module. Finally, the attentive visual knowledge features $\tilde{F}_v^g$, the attentive caption knowledge features $\tilde{F}_c^g$, and the features F^{vh} that are the output of the Multi-Modality Fusion Module, are fused by element-wise multiplication. The fused features are finally fed into the decoder to generate the corresponding answer.

3.1.4. Generative and discriminative decoder

The decoder in this task adopts the idea of multi-task learning. It is a combination of generative and discriminative decoders. In the discriminative decoder, the embedded candidate answers are ranked by the dot-product between the candidate answers and the fused features. Then a softmax function is used to produce the distribution of the candidate answers. The discriminative loss is the cross-entropy loss. For the generative decoder, two LSTMs initialized by the representation of encoder output are applied to predict the answers. Its corresponding loss is negative log-likelihood. During training, the generative loss and the discriminative loss are added.

3.2. Task 2: goal-oriented visual dialog: ‘Image Guessing’

In order to enrich the application scenarios of the structured knowledge and further demonstrate its effectiveness, we apply SKANet to the goal-oriented visual dialog (‘image guessing’). Specifically, a cooperative image guessing game is played by two agents(Q-BOT and A-BOT), which is shown in Fig. 3. In this subsection, we focus on the image guessing which is accomplished by Q-BOT. The agents in this game are mainly learned by reinforcement learning. The specific elements about reinforcement learning at round t are as follows:

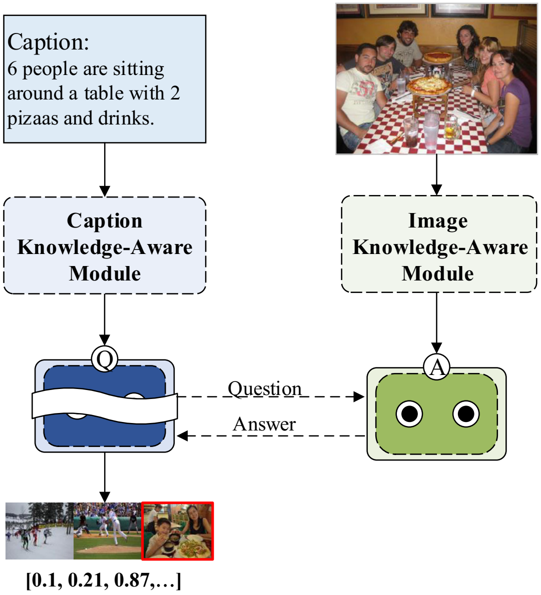

Caption:
6 people are sitting around a table with 2 pizaas and drinks.
Caption Knowledge-Aware Module
Image Knowledge-Aware Module
Q
A
Question
Answer
[0.1, 0.21, 0.87,…]

Fig. 3. The overview structure of SKANet for the ‘image guessing’ subtask. The two agents interact with each other on the condition of information asymmetry. A-BOT can see the image, and Q-BOT just knows the caption of the image. At the end of each round of dialog, Q-BOT completes the image retrieval from an image gallery based on the generated dialogs.

3.2.1. Agent and environment

‘Image guessing’ involves two agents including Q-BOT and A-BOT. The input image *I* is the environment.

3.2.2. Action

The action space of these two agents is sequences of English tokens. Q-BOT specifically generates the question q_t, and A-BOT produces the corresponding answer a_t. After receiving the answer a_t, Q-BOT predicts a supposed representation f_t of the unseen image so that it can complete the final guessing task.

3.2.3. State

Since Q-BOT can not see the image *I*, the state of Q-BOT S_t^Q consists of the dialog history $H_t^Q = \{(q_1, a_1), (q_2, a_2), \cdots, (q_{t-1}, a_{t-1})\}$, the caption *C*, and the caption knowledge features F_c^g, which are the outputs of the Caption Knowledge-Aware Module. The state of A-BOT S_t^A contains the dialog history H_t^A, the image features *V*, and the image knowledge features F_v^g.

3.2.4. Policy

For the generation of the dialog, the policies of Q-BOT and A-BOT are respectively $\pi_Q(q_t|s_t^Q)$ and $\pi_A(a_t|s_t^A)$. The policy networks of Q-BOT and A-BOT are shown as Fig. 4. For image guessing, a feature regression network (FRN) is designed for Q-BOT. FRN is responsible for generating the supposed representation $f_t = FRN(s_t^Q, q_t, a_t)$.

The policy network of Q-BOT contains four parts: Question Decoder, QA Pair Encoder, History Encoder, and FRN. Question Decoder firstly generates the current question q_t based on the state of the round $t-1$. Then Q-BOT receives the answer a_t generated by A-BOT and encodes the pair (q_t, a_t) by QA Pair Encoder. The encoded pair P_t^Q and the caption knowledge features F_c^g are input into the History Encoder to update S_{t-1}^Q to S_t^Q. Finally, FRN uses the output of the History Encoder to produce the supposed representation f_t and the reward r_t. In particular, Question Decoder, QA

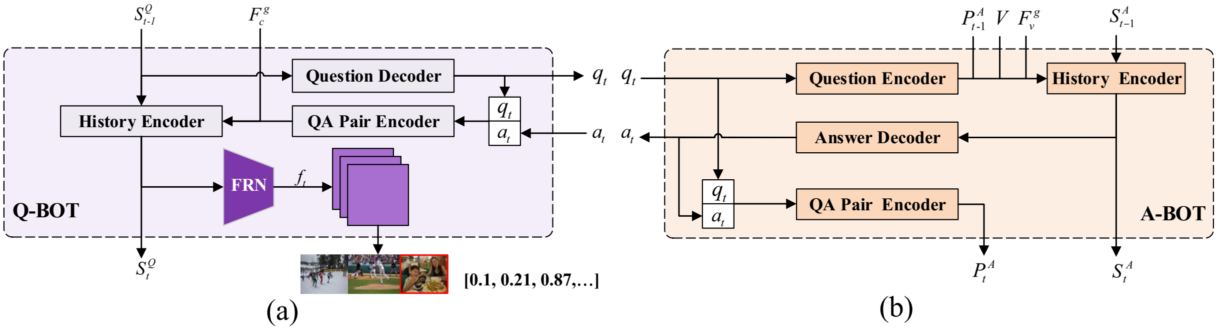


Fig. 4. The policy network of Q-BOT and A-BOT. At the round t, Q-BOT firstly generates q_t depending on the previous state S^Q_{t-1}. A-BOT decodes an answer a_t based on the current state S^A_t which is updated by the previous QA pair P^A_{t-1}, the image features V, and the image knowledge features F^g_v. Then Q-BOT updates its state S^Q_t by the current QA pairs P^Q_t and the caption knowledge features F^g_c. Finally, Q-BOT produces the representation f_t by the Feature Regression Network (FRN).

Pair Encoder, and History Encoder are all LSTM, and FRN is a simple fully-connected layer.

The policy network of A-BOT is composed of Question Encoder, History Encoder, Answer Decoder, and QA Pair Encoder. q_t from Q-BOT is encoded by Question Encoder. Then the encoded q_t, the previous encoded pair P^A_{t-1}, the image features V, and the image knowledge features F^g_v are utilized by history Encoder to update its state to S^A_t which is input into Answer Decoder to produce the answer a_t. QA Pair Encoder in A-BOT is the same as Q-BOT. Moreover, Question Encoder, History Encoder, Answer Decoder, and QA Pair Encoder are also achieved by LSTM.

3.2.5. Reward

We apply Euclidean distance between the supposed representation f_t and the ground truth image features f^{gt} to measure the reward $r_t = d\left(f_{t-1}, f^{gt}\right) - d\left(f_t, f^{gt}\right)$. The goal is to maximize the expected reward which is:

$$J\left(w^Q, w^A, w^{FRN}\right) = \underset{\pi_Q,\pi_A}{\mathbb{E}}\left[r_t\left(s^Q_t, (q_t, a_t, f_t)\right)\right]. \tag{6}$$

In particular, REINFORCE algorithm [54] is applied to train the agents.

3.2.6. Training strategy

Following [26], there are two training steps including supervised pretraining and curriculum learning [55]. Supervised pretraining aims to make Q-BOT and A-BOT learn the basic visual recognition ability and English question/answer generation ability. Curriculum learning is applied to fine-tune the pre-trained two agents. It helps Q-BOT and A-BOT produce successive and meaningful dialog. We firstly pre-train Q-BOT and A-BOT on VisDial v1.0 train split with an MLE loss. Then, supervised pretraining and reinforcement learning are recurrently applied to fine-tune the agents in a manner of curriculum. Specifically, we pre-train the agents for the first K rounds (K is initially set as 9), and then reinforcement learning is used to train the agents for the rest of 10-K rounds. K is reduced by 1 at a time, from 9 to 0.

4. Experiments

4.1. Dataset

All of our experiments are conducted on VisDial v0.9 and v1.0 datasets [16] collected from Amazon Mechanical Turk. The training and validation split of VisDial v0.9 dataset contain 82,783 and 40,504 images from COCO train 2014 and val 2014, respectively. There are fixed ten rounds of dialog per image. VisDial v1.0 dataset can be regarded as the extended version of VisDial v0.9. For VisDial v1.0, its training split is made up of the training and validation splits of VisDial v0.9. Its validation split and test split consist of 2,064 and 8,000 images from Flick, respectively. There is only one round dialog in the test split.

4.2. Evaluation metrics

We follow [16,26] to evaluate our method.

4.2.1. Task 1: conventional visual dialog

The accuracy of visual dialog is conducted by retrieving the ground-truth answer from the 100 candidate options. There are four kinds of evaluation metrics: 1) Mean Reciprocal Rank (MRR) of human response; 2) Recall@k (R@k); 3) Mean Rank of the human response; 4) Normalized Discounted Cumulative Gain (NDCG). The k in Recall@k is set as 1, 5, and 10. NDCG is a new evaluation metric proposed in VisDial v1.0. It penalizes the low-ranked but correct answers. The lower the better for Mean, and the others are the opposite.

4.2.2. Task 2: goal-oriented visual dialog: 'Image Guessing'

The evaluation metric for the generated dialogs is the same as the conventional visual dialog. The mean percentile rank is applied as the evaluation metric for the guessing ability. Specifically, a percentile rank of $K\%$ means that the ground truth image is closer to the predicted image than $K\%$ images in the test split. All of the test images are first sorted by the Euclidean distance between the predicted representation and the test images. And then, the ground truth image is ranked by the previous sorted test images.

4.3. Implementation details

The models in this paper are implemented by PyTorch 1.3 and Deep Graph Library (DGL) that is specific for graph neural network.

4.3.1. Task 1: conventional visual dialog

We mainly follow the code from [52]. The hidden state in BiLSTM in the preparation work and the graph features encoded by GCN are both 512-d. The batch size in the training phase is 32. We apply Adam as the optimizer and train the model by totally 6 epochs, including 2 warm-up epochs in which the learning rate is 1×10^{-5}. Then the learning rate is set as 1×10^{-3} in the following 4 epochs.

4.3.2. Task 2: goal-oriented visual dialog: 'Image Guessing'

The hidden state of each LSTM in the policy network of Q-BOT and A-BOT is 512-d. We pretrain the agents for 35 epochs on Vis-

Dial v1.0 dataset, and then fine-tune it by curriculum learning for 10 epochs.Similar to the previous work, the optimizer is also Adam, and the learning rate is set as 1×10^{-3}.

All of the experiments are conducted on an Nvidia Tesla V100 GPU with 32 GB memory.

4.4. Quantitative result

4.4.1. Task 1: conventional visual dialog

We compare SKANet with previous methods on VisDial v1.0 and 0.9 datasets. The methods contain LF [16], HRE [16], MN [16],VGNN [18], CorefNMN [17], RvA [40], DAN [19], FGA [13], DualVD [56], CAG [41]. LF directly concatenates the embedded features of the image, the question, and the history. HRE proposes a hierarchical recurrent encoder to embed the inputs. MN applies a memory network to store the dialog history, which is read by the encoded question and image. VGNN formalizes visual dialog as an inference of a graph model. CorefNMN, RvA and DAN separately uses neural module network, a recursive visual attention, and a dual attention network to resolve visual co-reference resolution. DualVD utilizes the dual encode theory to excavate the visual information. CAG builds a complex dynamic graph structure to explore the related context information. Almost all of these methods try to improve the quality of dialog from the perspective of data itself or data interaction. In addition to these, our model applies the external commonsense knowledge to enhance the reasoning ability. Table 1 shows the quantitative results of our method.

For VisDial v0.9, SKANet almost outperforms all comparison methods on MRR, Recall@K, and Mean. Compared with the typical attention-based methods RvA and DAN, our method improves the performance of MRR by more than 1.1%. For R@10 metric, it achieves about 0.5% improvement. VisDial v0.9 is sparser compared with VisDial v1.0. In addition to the COCO dataset, VisDial v1.0 also applies the Flick dataset, which makes the data richer and more difficult. Meanwhile, the external knowledge is more effective for the diverse data, so our method is a little inferior to CAN on VisDial v0.9.

For VisDial v1.0, our method achieves the state-of-art results. Since NDCG quantifies whether all the correct answers are ranked high, it is somewhat mutually exclusive with other metrics. However, our method outperforms other models on all evaluation metrics. Compared with CAG, our SKANet model achieves more than 1.3% improvement on MRR and NDCG. Mean metric is boosted from 4.11 to 3.98. Moreover, we improve the NDCG by about 1.3%. They all show the superiority of our method for the conventional visual dialog task.

4.4.2. Task 2: goal-oriented visual dialog: ‘Image Guessing’

We select [26] as the baseline of the cooperative image guessing game. Fig. 5 represents the image guessing performance of Q-BOT. As the dialog round increases, the percentile of SL-Pretrain is 91.59%. Under the impact of the Knowledge-Aware Module (KAM), the percentile can reach 92.63%. After the fine-tune by reinforcement learning in a manner of curriculum, the percentile without the KAM is about 95%. Finally, KAM boosts it by about 0.8%.

Moreover, we also validate the answers during the guessing process on VisDial v1.0 validation split. Different from the conventional visual dialog, the questions in ‘image guessing’ are generated by the agent Q-BOT. Many repetitive dialogs are inevitable, so its experimental results drop a lot. However, KAM still improves the quality of the answers obviously. The specific results are shown as Table 2. If we only pretrain the model by supervised learning, KAM boosts the R@5 metric from 55.05 to 55.75. And the Mean is improved to 16.69. For the full training process, all evaluation metrics are improved more significantly. The MRR metric is improved from 46.43 to 47.06 under the impact of KAM. R@10 increases by about 0.5%, and the Mean drops to 18.88. They further validate the effectiveness of our method.

4.5. Ablation study

In Table 3, an ablation study about SKANet is performed on VisDial v1.0 validation split to further verify the validity of the main module. We primarily ablate the Knowledge-Aware Module (KAM). The baseline only retains MMFM (the Multi-Modality Fusion Module). The model with KAM-ConceptNet means that it ablates the ConceptNet in KAM. We directly construct fully connected graph over the detected concepts from the image and the caption. Compared with the baseline, its results show little improvement. However, the model with KAM obviously improves the performance on all the evaluation metrics. For example, the model with KAM achieves 1.3% on MRR and boosts Recall@K by about 1.3%on average. It proves that the commonsense knowledge virtually improves the quality of the dialogue. Finally, ML (multi-task learning) is applied to our method, as mentioned in Section 3.1.4. It further improves the results. Moreover, the visualization results are shown in the next section to explain our model intuitively.

4.6. Qualitative result

The qualitative results in Fig. 6, Fig. 7 and Fig. 8 further demonstrate the effectiveness of our SKANet model.

Table 1
Quantitative results of our model on VisDial V1.0 test split.

	VisDial v1.0						VisDial v0.9				
Model	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓	NDCG ↑	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓
LF [16]	55.42	40.95	72.45	82.83	5.95	45.31	58.07	43.82	74.68	84.07	5.78
HRE [16]	54.16	39.93	70.47	81.50	6.41	45.46	58.46	44.67	74.50	84.22	5.72
MN [16]	55.49	40.98	72.30	83.30	5.92	47.50	59.65	45.55	76.22	85.37	5.46
VGNN [18]	61.37	47.33	77.98	87.83	4.57	52.82	62.85	48.95	79.65	88.36	4.57
CorefNMN [17]	61.50	47.55	78.10	88.80	4.40	54.70	64.10	50.92	80.18	88.81	4.45
RvA [40]	63.03	49.03	80.40	89.83	4.18	55.59	66.34	52.71	82.97	90.73	3.93
DAN [19]	63.20	49.63	79.75	89.35	4.30	57.59	66.38	53.33	82.42	90.38	4.04
FGA [13]	63.70	49.58	80.98	88.55	4.51	52.10	67.12	54.02	83.21	90.47	4.08
DualVD [56]	63.23	49.25	80.23	89.70	4.11	56.32	62.94	48.64	80.89	89.94	4.17
CAG [41]	63.49	49.85	80.63	90.15	4.11	56.64	**67.56**	54.64	**83.72**	**91.48**	**3.75**
SKANet (ours)	**64.79**	**51.25**	**81.18**	**90.43**	**3.98**	**57.97**	67.52	**54.64**	83.52	91.25	3.78

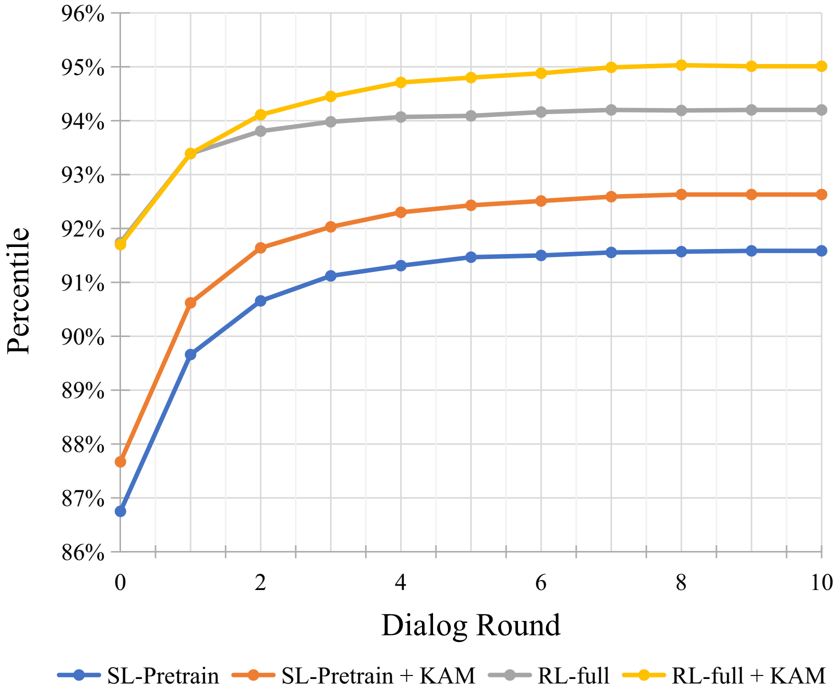


Fig. 5. Image Guessing Performance of Q-BOT. SL and RL separately mean the supervised learning and the reinforcement learning. KAM represents the Knowledge-Aware Module.

Table 2
The answer evaluation of goal-oriented visual dialog on VisDial v1.0 val split.

Model	Epoch	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓
SL-Pretrain	35	45.53	36.17	55.05	61.41	19.79
SL-Pretrain + Knowledge Module	35	45.81	35.50	55.75	61.91	19.69
RL-full	10	46.43	36.33	56.27	62.81	19.27
RL-full + Knowledge Module	10	47.06	37.19	56.52	63.31	18.88

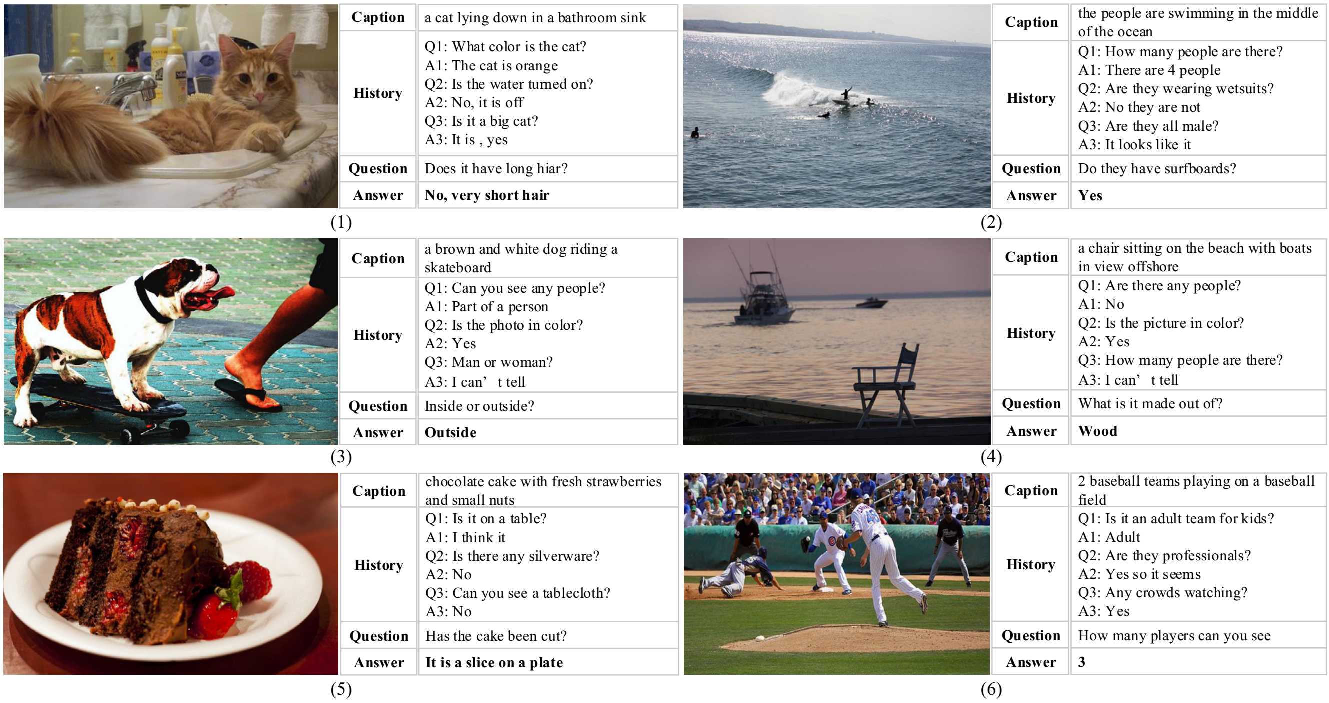
Table 3
Ablation study of our model on VisDial v1.0 validation split. MMFM and KAM separately denote the Multi-Modality Fusion Module and the Knowledge-Aware Module. ML represents the multi-task learning. Without ML, the model only adapts discriminative decoder.

MMFM	KAM-ConceptNet	KAM	ML	MRR ↑	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓
√				64.15	50.59	80.73	89.72	4.14
√	√			64.19	50.65	80.71	89.84	4.13
√		√		65.46	52.08	82.14	90.52	3.92
√		√	√	**65.60**	**52.19**	**82.34**	**91.12**	**3.86**

4.6.1. Task 1: conventional visual dialog

Fig. 6 shows that our model can be capable of a wide range of scenarios from simple to complex. Fig. 6(1) and Fig. 6(2) are the examples with simple scenes. They only needs the visual features or the textual history features to answer the questions. For example, the surfboard is detected in the image of Fig. 6(2), then the question "Do they have surfboards?" can be answered easily. Fig. 6(3), Fig. 6(4) and Fig. 6(5) represent the examples with more complex scenarios. The information from the image and the caption is insufficient to answer some questions. However, our model can solve them by the external knowledge. For example, the chair in Fig. 6(4) is made out of wood, but the visual features and the textual features are hard to directly answer it. The answer must be resolved by the latent knowledge about the chair and the beach. Meanwhile, there are actually five players in the image in Fig. 6 (6), but the answer is "three". Our model cannot well handle the complicated counting problem. The richer visual representation may need to be excavated, and the deep reasoning ability also needs to be improved.

In addition, we validate the effectiveness of the commonsense knowledge in detail. Two examples with specific external knowledge are given in Fig. 7. The first example with complex sense demands to answer the material of the table in the image. Whereas there are lots of things on it, it is hard to directly identify its material. Then much related external knowledge about the image and its caption is extracted, such as "Bed", "House", "Wood", and "Apple". The "Wood" can be used to accomplish the question answering. If the key component KAM which applies the external knowledge is not uesd, the answer is "I can't tell". The objects in the second example are relatively simple. While the question about the kitchen is difficult because the image is full of foods. The visual representation and its caption cannot infer the information about the kitchen. Without the KAM, "I can't tell" is produced again. Then the drawn common sense from ConceptNet, including "Table",



(1)

Caption	a cat lying down in a bathroom sink
History	Q1: What color is the cat? A1: The cat is orange Q2: Is the water turned on? A2: No, it is off Q3: Is it a big cat? A3: It is , yes
Question	Does it have long hiar?
Answer	**No, very short hair**

(2)

Caption	the people are swimming in the middle of the ocean
History	Q1: How many people are there? A1: There are 4 people Q2: Are they wearing wetsuits? A2: No they are not Q3: Are they all male? A3: It looks like it
Question	Do they have surfboards?
Answer	**Yes**

(3)

Caption	a brown and white dog riding a skateboard
History	Q1: Can you see any people? A1: Part of a person Q2: Is the photo in color? A2: Yes Q3: Man or woman? A3: I can’ t tell
Question	Inside or outside?
Answer	**Outside**

(4)

Caption	a chair sitting on the beach with boats in view offshore
History	Q1: Are there any people? A1: No Q2: Is the picture in color? A2: Yes Q3: How many people are there? A3: I can’ t tell
Question	What is it made out of?
Answer	**Wood**

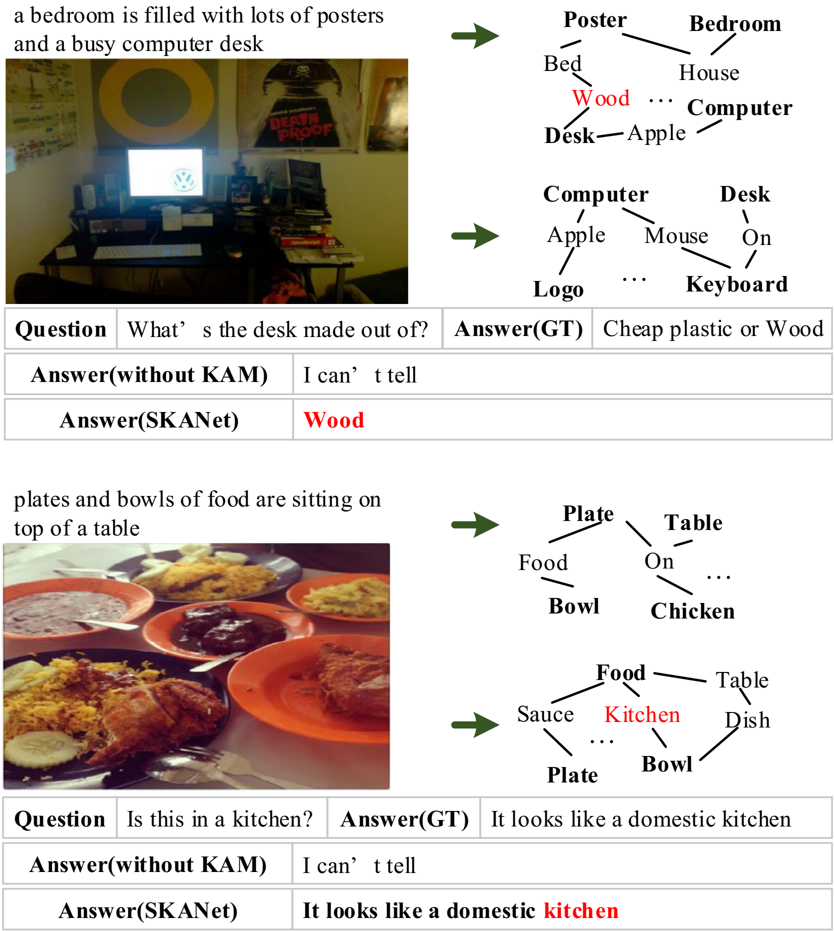
(5)

Caption	chocolate cake with fresh strawberries and small nuts
History	Q1: Is it on a table? A1: I think it Q2: Is there any silverware? A2: No Q3: Can you see a tablecloth? A3: No
Question	Has the cake been cut?
Answer	**It is a slice on a plate**

(6)

Caption	2 baseball teams playing on a baseball field
History	Q1: Is it an adult team for kids? A1: Adult Q2: Are they professionals? A2: Yes so it seems Q3: Any crowds watching? A3: Yes
Question	How many players can you see
Answer	**3**

Fig. 6. Qualitative results of our model. Our model is capable of a wide range of scenarios from simple to complex.



a bedroom is filled with lots of posters and a busy computer desk

Question	What’ s the desk made out of?	**Answer(GT)**	Cheap plastic or Wood
Answer(without KAM)	I can’ t tell		
Answer(SKANet)	**Wood**		

plates and bowls of food are sitting on top of a table

Question	Is this in a kitchen?	**Answer(GT)**	It looks like a domestic kitchen
Answer(without KAM)	I can’ t tell		
Answer(SKANet)	**It looks like a domestic kitchen**		

Fig. 7. Visualization of two examples with the commonsense knowledge.

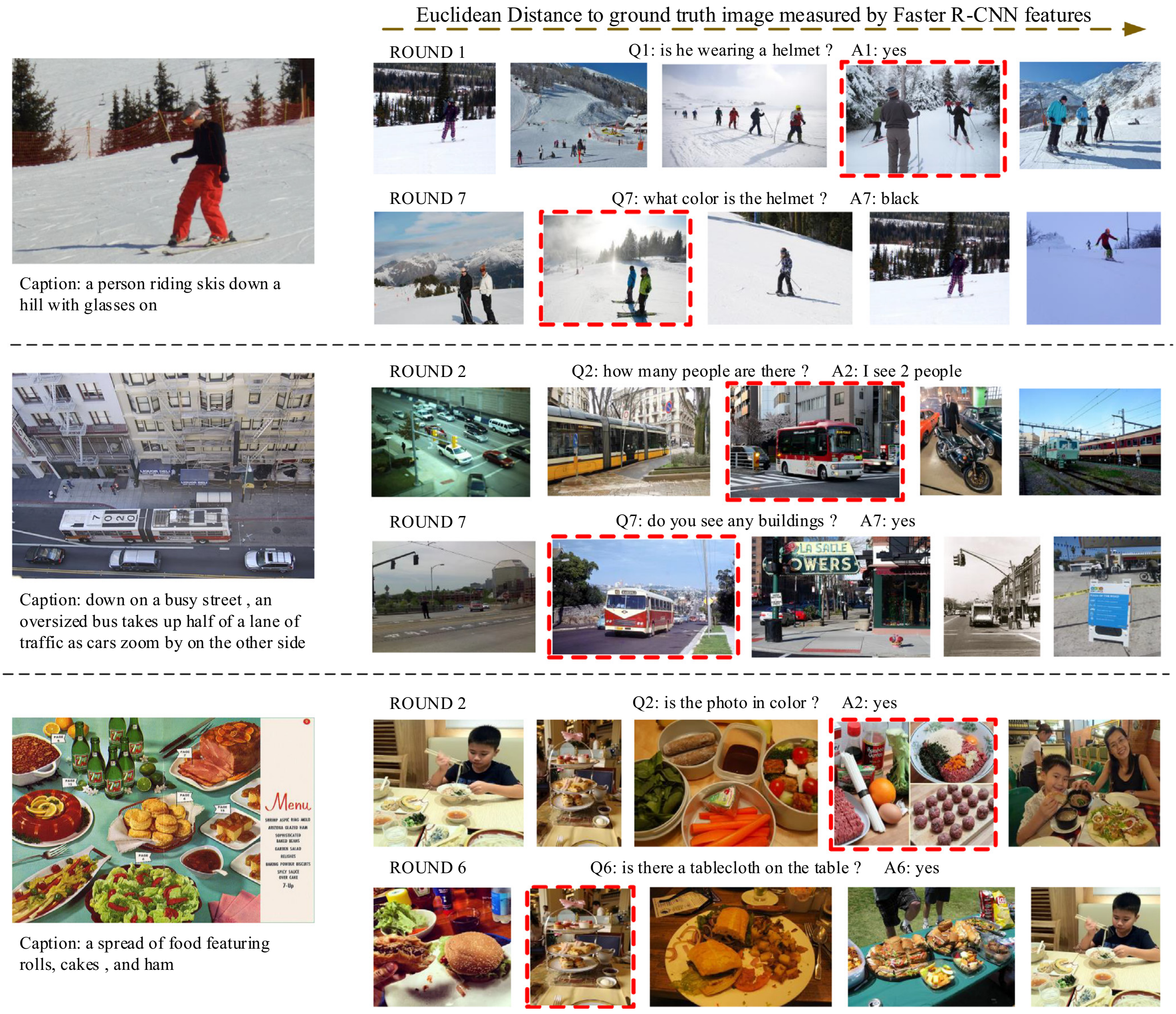

Euclidean Distance to ground truth image measured by Faster R-CNN features
ROUND 1
Q1: is he wearing a helmet ? A1: yes
ROUND 7
Q7: what color is the helmet ? A7: black
Caption: a person riding skis down a hill with glasses on
ROUND 2
Q2: how many people are there ? A2: I see 2 people
ROUND 7
Q7: do you see any buildings ? A7: yes
Caption: down on a busy street , an oversized bus takes up half of a lane of traffic as cars zoom by on the other side
ROUND 2
Q2: is the photo in color ? A2: yes
ROUND 6
Q6: is there a tablecloth on the table ? A6: yes
Caption: a spread of food featuring rolls, cakes , and ham
LA SALLE
OWERS
Menu

Fig. 8. The guessing results of our model on VisDial v1.0 validation split. The images in the red bounding box are the images guessed by Q-BOT.

"Dish", and "Kitchen", completes the reasoning and gives a correct answer. These examples intuitively show the importance of external knowledge for visual dialog.

4.6.2. Task 2: goal-oriented visual dialog: 'Image Guessing'

The qualitative results about the 'image guessing' task are shown in Fig. 8. The images in the left column are from the validation split of VisDial v1.0 dataset, and they are sorted by Euclidean distance to the ground truth image in the Faster R-CNN feature space. The images in the red bounding box are the results predicted by Q-BOT. In the first example of Fig. 8, the retrieved images in round 1 and round 7 are quite similar to the images seen by A-BOT from the perspective of human cognition. It validates the effectiveness of our proposed method.

5. Conclusion

In this paper, we learn the external commonsense knowledge for visual dialog and propose a novel pipeline named SKANet to handle the complex scenarios. We extract the related knowledge from ConceptNet through the concept set of the input image and its corresponding caption. Furthermore, GCN is applied to extract structured graph features. These features help the agent enhance the ability of reasoning rationality. Moreover, in order to demonstrate the availability of the main components of SKANet, we apply this model to the conventional visual dialog and a goal-oriented visual dialog named 'image guessing'. The experiments conducted on VisDial v0.9 and v1.0 datasets validate the effectiveness of our proposed method.

CRediT authorship contribution statement

Lei Zhao: Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Haonan Zhang:** Resources, Writing - review & editing. **Xiangpeng Li:** Investigation, Writing - review & editing. **Sen Yang:** Data curation, Validation. **Yuanfeng Song:** Funding acquisition, Investigation, Software, Resources, Visualization, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

aaa

References

[1] 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) 6077–6086.

[2] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, H.T. Shen, From deterministic to generative: Multimodal stochastic rnns for video captioning, IEEE Trans. Neural Networks Learn. Syst. 30 (10) (2019) 3047–3058.

[3] J. Yang, Y. Sun, J. Liang, B. Ren, S. Lai, Image captioning by incorporating affective concepts learned from both visual and textual components, Neurocomputing 328 (2019) 56–68.

[4] L. Gao, X. Li, J. Song, H.T. Shen, Hierarchical lstms with adaptive attention for visual captioning, IEEE Trans. Pattern Anal. Mach. Intell. 42 (5) (2020) 1112–1131.

[5] X. Yang, J. Dong, Y. Cao, X. Wang, M. Wang, T.-S. Chua, Tree-augmented cross-modal encoding for complex-query video retrieval, in, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1339–1348.

[6] X. Yang, F. Feng, W. Ji, M. Wang, T.-S. Chua, Deconfounded video moment retrieval with causal intervention, in, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1–10.

[7] J. Dong, X. Li, C. Xu, X. Yang, G. Yang, X. Wang, M. Wang, Dual encoding for video retrieval by text, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[8] Y. Qiao, Z. Yu, J. Liu, Rankvqa: Answer re-ranking for visual question answering, in: ICME, 2020, pp. 1–6.

[9] X. Li, J. Song, L. Gao, X. Liu, W. Huang, X. He, C. Gan, Beyond rnns: Positional self-attention with co-attention for video question answering, in, AAAI (2019) 8658–8665.

[10] X. Li, L. Gao, X. Wang, W. Liu, X. Xu, H.T. Shen, J. Song, Learnable aggregating net with diversity learning for video question answering, in: ACM MM, ACM, 2019, pp. 1166–1174.

[11] J. Hong, J. Fu, Y. Uh, T. Mei, H. Byun, Exploiting hierarchical visual features for visual question answering, Neurocomputing 351 (2019) 187–195.

[12] Q. Wang, Y. Han, Visual dialog with targeted objects, in, ICME (2019) 1564–1569.

[13] I. Schwartz, S. Yu, T. Hazan, A.G. Schwing, Factor graph attention, CVPR (2019) 2039–2048.

[14] Y. Chang, W. Peng, Learning goal-oriented visual dialog agents: Imitating and surpassing analytic experts, ICME (2019) 520–525.

[15] D. Guo, C. Xu, D. Tao, Image-question-answer synergistic network for visual dialog, CVPR (2019) 10434–10443.

[16] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J.M.F. Moura, D. Parikh, D. Batra, Visual dialog, in: CVPR, 2017, pp. 1080–1089.

[17] S. Kottur, J.M.F. Moura, D. Parikh, D. Batra, M. Rohrbach, Visual coreference resolution in visual dialog using neural module networks, in: ECCV, Vol. 11219, 2018, pp. 160–178.

[18] Z. Zheng, W. Wang, S. Qi, S. Zhu, Reasoning visual dialogs with structural and partial observations, in, CVPR (2019) 6669–6678.

[19] G. Kang, J. Lim, B. Zhang, Dual attention networks for visual reference resolution in visual dialog, in, EMNLP-IJCNLP (2019) 2024–2033.

[20] X. Yang, X. Liu, M. Jian, X. Gao, M. Wang, Weakly-supervised video object grounding by exploring spatio-temporal contexts, in: in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1939–1947.

[21] X. Liu, X. Yang, M. Wang, R. Hong, Deep neighborhood component analysis for visual similarity modeling, ACM Transactions on Intelligent Systems and Technology (TIST) 11 (3) (2020) 1–15.

[22] J. Xiao, X. Shang, X. Yang, S. Tang, T.-S. Chua, Visual relation grounding in videos, in, European Conference on Computer Vision (2020) 447–464.

[23] Y. Li, X. Yang, X. Shang, T.-S. Chua, Interventional video relation detection, in, ACM International Conference on Multimedia (2021).

[24] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, A.C. Courville, Guesswhat?! visual object discovery through multi-modal dialogue, in, CVPR (2017) 4466–4475.

[25] B. Zhuang, Q. Wu, C. Shen, I.D. Reid, A. van den Hengel, Parallel attention: A unified framework for visual object discovery through dialogs and queries, in: CVPR, 2018, pp. 4252–4261.

[26] A. Das, S. Kottur, J.M.F. Moura, S. Lee, D. Batra, Learning cooperative visual dialog agents with deep reinforcement learning, in: ICCV, 2017, pp. 2970–2979.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in, NeurIPS (2017) 5998–6008.

[28] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, in: ICLR, OpenReview.net, 2018.

[29] S. Ren, K. He, R.B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149.

[30] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D.A. Shamma, M.S. Bernstein, L. Fei-Fei, Visual genome: Connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (1) (2017) 32–73.

[31] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in, AAAI (2017) 4444–4451.

[32] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: ICLR, OpenReview.net, 2017.

[33] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C.L. Zitnick, D. Parikh, D. Batra, VQA: visual question answering - www.visualqa.org, Int. J. Comput. Vis. 123 (1) (2017) 4–31.

[34] L. Ma, Z. Lu, H. Li, Learning to answer questions from image using convolutional neural network, in, AAAI (2016) 3567–3573.

[35] K. Saito, A. Shin, Y. Ushiku, T. Harada, Dualnet: Domain-invariant network for visual question answering, in, ICME (2017) 829–834.

[36] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in, NeurIPS (2016) 289–297.

[37] 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. 1839–1848.

[38] Z. Yu, J. Yu, Y. Cui, D. Tao, Q. Tian, Deep modular co-attention networks for visual question answering, in, CVPR (2019) 6281–6290.

[39] D. Nguyen, T. Okatani, Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering, in, CVPR (2018) 6087–6096.

[40] Y. Niu, H. Zhang, M. Zhang, J. Zhang, Z. Lu, J. Wen, Recursive visual attention in visual dialog, in, CVPR (2019) 6679–6688.

[41] D. Guo, H. Wang, H. Zhang, Z. Zha, M. Wang, Iterative context-aware graph inference for visual dialog, in, CVPR (2020) 10052–10061.

[42] U. Jain, S. Lazebnik, A.G. Schwing, Two can play this game: Visual dialog with discriminative question generation and answering, in, CVPR (2018) 5754–5763.

[43] D. Massiceti, N. Siddharth, P.K. Dokania, P.H.S. Torr, Flipdial: A generative model for two-way visual dialogue, in: CVPR, 2018, pp. 6097–6105.

[44] R. Shekhar, T. Baumgärtner, A. Venkatesh, E. Bruni, R. Bernardi, R. Fernández, Ask no more: Deciding when to guess in referential visual dialogue, in, COLING (2018) 1218–1233.

[45] L. Zhao, X. Lyu, J. Song, L. Gao, Guesswhich? visual dialog with attentive memory network, Pattern Recognition 114 (2021) 107823.

[46] A. Agarwal, S. Gurumurthy, V. Sharma, M. Lewis, K.P. Sycara, Community regularization of visually-grounded dialog, in, AAMAS (2019) 1042–1050.

[47] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z.G. Ives, Dbpedia: A nucleus for a web of open data, in: The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11–15, 2007, Vol. 4825, 2007, pp. 722–735.

[48] M. Sap, R.L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N.A. Smith, Y. Choi, ATOMIC: an atlas of machine commonsense for if-then reasoning, in, AAAI (2019) 3027–3035.

[49] S. Bhakthavatsalam, C. Anastasiades, P. Clark, Genericskb: A knowledge base of generic statements, arXiv preprint arXiv:2005.00660.

[50] S. Shah, A. Mishra, N. Yadati, P.P. Talukdar, KVQA: knowledge-aware visual question answering, in: AAAI, 2019, pp. 8876–8884.

[51] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: EMNLP, 2014, pp. 1532–1543.

[52] B.Y. Lin, X. Chen, J. Chen, X. Ren, Kagnet: Knowledge-aware graph networks for commonsense reasoning, in: EMNLP-IJCNLP, 2019, pp. 2829–2839.

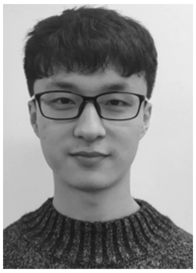
[53] Z. Wang, J. Zhang, J. Feng, Z. Chen, Knowledge graph embedding by translating on hyperplanes, in, AAAI (2014) 1112–1119.

[54] R.J. Williams, Simple statistical gradient-following algorithms for connectionist reinforcement learning, Mach. Learn. 8 (1992) 229–256.

[55] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: ICML, Vol. 382, 2009, pp. 41–48.

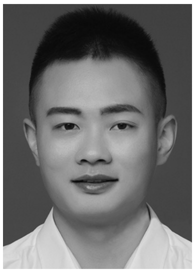
[56] X. Jiang, J. Yu, Z. Qin, Y. Zhuang, X. Zhang, Y. Hu, Q. Wu, Dualvd: An adaptive dual encoding model for deep visual understanding in visual dialogue, in, AAAI (2020) 11125–11132.

Lei Zhao is a PH.D. student in the School of Computer Science and Engineering, University of Electronic Science and Technology of China. Currently, he is working on visual dialog, and video object segmentation.

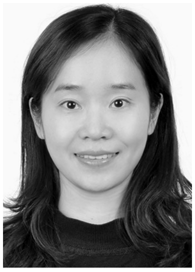


Haonan Zhang is a M.S. student in the School of Computer Science and Engineering, University of Electronic Science and Technology of China. Currently, he is working on visual dialog and video question answering.

Xiangpeng Li is a PH.D. student in the School of Computer Science and Engineering, University of Electronic Science and Technology of China. Currently, he is working on image understanding, image/video captioning and visual dialog.



Sen Yang is a M.S. student in the School of Computer Science and Engineering, University of Electronic Science and Technology of China. Currently, he is working on densepose and visual dialog.



Yuanfeng Song is currently a lecturer in Information Center, University of Electronic Science and Technology of China, Chengdu, China. She received her master degree in computer system structure from University of Electronic Science and Technology of China. Her current research interests include machine learning and data mining.