Incorporating External Knowledge to Answer Open-Domain Visual Questions with Dynamic Memory Networks

Abstract

Visual Question Answering (VQA) has attracted much attention since it offers insight into the relationships between the multi-modal analysis of images and natural language. Most of the current algorithms are incapable of answering open-domain questions that require to perform reasoning beyond the image contents. To address this issue, we propose a novel framework which endows the model capabilities in answering more complex questions by leveraging massive external knowledge with dynamic memory networks. Specifically, the questions along with the corresponding images trigger a process to retrieve the relevant information in external knowledge bases, which are embedded into a continuous vector space by preserving the entity-relation structures. Afterwards, we employ dynamic memory networks to attend to the large body of facts in the knowledge graph and images, and then perform reasoning over these facts to generate corresponding answers. Extensive experiments demonstrate that our model not only achieves the state-of-the-art performance in the visual question answering task, but can also answer open-domain questions effectively by leveraging the external knowledge.

Introduction

Visual Question Answering (VQA) is a ladder towards a better understanding of the visual world, which pushes forward the boundaries of both computer vision and natural language processing. A system in VQA tasks is given a text-based question about an image, which is expected to generate a correct answer corresponding to the question. In general, VQA is a kind of Visual Turing Test, which rigorously assesses whether a system is able to achieve human-level semantic analysis of images BIBREF0 , BIBREF1 . A system could solve most of the tasks in computer vision if it performs as well as or better

than humans in VQA. In this case, it has garnered increasing attentions due to its numerous potential applications BIBREF2 , such as providing a more natural way to improve human-computer interaction, enabling the visually impaired individuals to get information about images, etc.

To fulfill VQA tasks, it requires to endow the responder to understand intention of the question, reason over visual elements of the image, and sometimes have general knowledge about the world. Most of the present methods solve VQA by jointly learning interactions and performing inference over the question and image contents based on the recent success of deep learning BIBREF3 , BIBREF2 , BIBREF4 , BIBREF5 , BIBREF6 , which can be further improved by introducing the attention mechanisms BIBREF7 , BIBREF8 , BIBREF9 , BIBREF10 , BIBREF11 , BIBREF12 . However, most of questions in the current VQA dataset are quite simple, which are answerable by analyzing the question and image alone BIBREF2 , BIBREF13 . It can be debated whether the system can answer questions that require prior knowledge ranging common sense to subject-specific and even expert-level knowledge. It is attractive to develop methods that are capable of deeper image understanding by answering open-domain questions BIBREF13 , which requires the system to have the mechanisms in connecting VQA with structured knowledge, as is shown in Fig. 1 . Some efforts have been made in this direction, but most of them can only handle a limited number of predefined types of questions BIBREF14 , BIBREF15 .

Different from the text-based QA problem, it is unfavourable to conduct the open-domain VQA based on the knowledge-based reasoning, since it is inevitably incomplete to describe an image with structured forms BIBREF16 . The recent availability of large training datasets BIBREF13 makes it feasible to train a complex model in an end-to-end fashion by leveraging the recent advances in deep neural networks (DNN) BIBREF2 , BIBREF5 , BIBREF7 , BIBREF10 , BIBREF12 . Nevertheless, it is non-trivial to integrate knowledge into DNN-based methods, since the knowledge are usually represented in a symbol-based or graph-based manner (e.g., Freebase BIBREF17 , DBPedia BIBREF18), which is intrinsically different from the DNN-based features. A few attempts are made in this direction BIBREF19 ,

but it may involve much irrelevant information and fail to implement multi-hop reasoning over several facts.

The memory networks BIBREF20 , BIBREF21 , BIBREF22 offer an opportunity to address these challenges by reading from and writing to the external memory module, which is modeled by the actions of neural networks. Recently, it has demonstrated the state-of-the-art performance in numerous NLP applications, including the reading comprehension BIBREF23 and textual question answering BIBREF24 , BIBREF22 . Some seminal efforts are also made to implement VQA based on dynamic memory networks BIBREF25 , but it does not involve the mechanism to incorporate the external knowledge, making it incapable of answering open-domain visual questions. Nevertheless, the attractive characteristics motivate us to leverage the memory structures to encode the large-scale structured knowledge and fuse it with the image features, which offers an approach to answer open domain visual questions.

Our Proposal

To address the aforementioned issues, we propose a novel Knowledge-incorporated Dynamic Memory Network framework (KDMN), which allows to introduce the massive external knowledge to answer open-domain visual questions by exploiting the dynamic memory network. It endows a system with an capability to answer a broad class of open-domain questions by reasoning over the image content incorporating the massive knowledge, which is conducted by the memory structures.

Different from most of existing techniques that focus on answering visual questions solely on the image content, we propose to address a more challenging scenario which requires to implement reasoning beyond the image content. The DNN-based approaches BIBREF2 , BIBREF5 , BIBREF7 are therefore not sufficient, since they can only capture information present in the training images. Recent advances

witness several attempts to link the knowledge to VQA methods BIBREF14 , BIBREF15 , which make use of structured knowledge graphs and reason about an image on the supporting facts. Most of these algorithms first extract the visual concepts from a given image, and implement reasoning over the structured knowledge bases explicitly. However, it is non-trivial to extract sufficient visual attributes, since an image lacks the structure and grammatical rules as language. To address this issue, we propose to retrieve a bath of candidate knowledge corresponding to the given image and related questions, and feed them to the deep neural network implicitly. The proposed approach provides a general pipeline that simultaneously preserves the advantages of DNN-based approaches BIBREF2 , BIBREF5 , BIBREF7 and knowledge-based techniques BIBREF14 , BIBREF15 .

In general, the underlying symbolic nature of a Knowledge Graph (KG) makes it difficult to integrate with DNNs. The usual knowledge graph embedding models such as TransE BIBREF26 focus on link prediction, which is different from VQA task aiming to fuse knowledge. To tackle this issue, we propose to embed the entities and relations of a KG into a continuous vector space, such that the factual knowledge can be used in a more simple manner. Each knowledge triple is treated as a three-word SVO $(subject, verb, object)$ phrase, and embedded into a feature space by feeding its word-embedding through an RNN architecture. In this case, the proposed knowledge embedding feature shares a common space with other textual elements (questions and answers), which provides an additional advantage to integrate them more easily.

Once the massive external knowledge is integrated into the model, it is imperative to provide a flexible mechanism to store a richer representation. The memory network, which contains scalable memory with a learning component to read from and write to it, allows complex reasoning by modeling interaction between multiple parts of the data BIBREF20 , BIBREF25 . In this paper, we adopt the most recent advance of Improved Dynamic Memory Networks (DMN+) BIBREF25 to implement the complex reasoning over several facts. Our model provides a mechanism to attend to candidate knowledge

embedding in an iterative manner, and fuse it with the multi-modal data including image, text and knowledge triples in the memory component. The memory vector therefore memorizes useful knowledge to facilitate the prediction of the final answer. Compared with the DMN+ BIBREF25 , we introduce the external knowledge into the memory network, and endows the system an ability to answer open-domain question accordingly.

To summarize, our framework is capable of conducting the multi-modal data reasoning including the image content and external knowledge, such that the system is endowed with a more general capability of image interpretation. Our main contributions are as follows:

To our best knowledge, this is the first attempt to integrating the external knowledge and image representation with a memory mechanism, such that the open-domain visual question answering can be conducted effectively with the massive knowledge appropriately harnessed;

We propose a novel structure-preserved method to embed the knowledge triples into a common space with other textual data, making it flexible to integrate different modalities of data in an implicit manner such as image, text and knowledge triples;

We propose to exploit the dynamic memory network to implement multi-hop reasonings, which has a capability to automatically retrieve the relevant information in the knowledge bases and infer the most probable answers accordingly.

Overview

In this section, we outline our model to implement the open-domain visual question answering. In order to conduct the task, we propose to incorporate the image content and external knowledge by exploiting the

most recent advance of dynamic memory network BIBREF22 , BIBREF25 , yielding three main modules in Fig. 2 . The system is therefore endowed with an ability to answer arbitrary questions corresponding to a specific image.

Considering of the fact that most of existing VQA datasets include a minority of questions that require prior knowledge, the performance therefore cannot reflect the particular capabilities. We automatically produce a collection of more challenging question-answer pairs, which require complex reasoning beyond the image contents by incorporating the external knowledge. We hope that it can serve as a benchmark for evaluating the capability of various VQA models on the open-domain scenarios .

Given an image, we apply the Fast-RCNN BIBREF27 to detect the visual objects of the input image, and extract keywords of the corresponding questions with syntax analysis. Based on these information, we propose to learn a mechanism to retrieve the candidate knowledge by querying the large-scale knowledge graph, yielding a subgraph of relevant knowledge to facilitate the question answering. During the past years, a substantial amount of large-scale knowledge bases have been developed, which store common sense and factual knowledge in a machine readable fashion. In general, each piece of structured knowledge is represented as a triple $(subject, rel, object)$ with $subject$ and $object$ being two entities or concepts, and rel corresponding to the specific relationship between them. In this paper, we adopt external knowledge mined from ConceptNet BIBREF28 , an open multilingual knowledge graph containing common-sense relationships between daily words, to aid the reasoning of open-domain VQA.

Our VQA model provides a novel mechanism to integrate image information with that extracted from the ConceptNet within a dynamic memory network. In general, it is non-trivial to integrate the structured knowledge with the DNN features due to their different modalities. To address this issue, we embed the entities and relations of the subgraph into a continuous vector space, which preserves the inherent structure of the KGs. The feature embedding provides a convenience to fuse with the image

representation in a dynamic memory network, which builds on the attention mechanism and the memory update mechanism. The attention mechanism is responsible to produce the contextual vector with relevance inferred by the question and previous memory status. The memory update mechanism then renews the memory status based on the contextual vector, which can memorize useful information for predicting the final answer. The novelty lies the fact that these disparate forms of information are embedded into a common space based on memory network, which facilities the subsequent answer reasoning.

Finally, we generate a predicted answer by reasoning over the facts in the memory along with the image contents. In this paper, we focus on the task of multi-choice setting, where several multi-choice candidate answers are provided along with a question and a corresponding image. For each question, we treat every multi-choice answer as input, and predict whether the image-question-answer triplet is correct. The proposed model tries to choose one candidate answer with the highest probability by inferring the cross entropy error on the answers through the entire network.

Answer Open-Domain Visual Questions

In this section, we elaborate on the details and formulations of our proposed model for answering open-domain visual questions. We first retrieve an appropriate amount of candidate knowledge from the large-scale ConceptNet by analyzing the image content and the corresponding questions; afterward, we propose a novel framework based on dynamic memory network to embed these symbolic knowledge triples into a continuous vector space and store it in a memory bank; finally, we exploit these information to implement the open-domain VQA by fusing the knowledge with image representation.

Candidate Knowledge Retrieval

In order to answer the open-domain visual questions, we should sometime access information not present in the image by retrieving the candidate knowledge in the KBs. A desirable knowledge retrieval should include most of the useful information while ignore the irrelevant ones, which is essential to avoid model misleading and reduce the computation cost. To this end, we take the following three principles in consideration as (1) entities appeared in images and questions (key entities) are critical; (2) the importance of entities that have direct or indirect links to key entities decays as the number of link hops increases; (3) edges between these entities are potentially useful knowledge.

Following these principles, we propose a three-step procedure to retrieve that candidate knowledge that are relevant to the context of images and questions. The retrieval procedure pays more attention on graph nodes that are linked to semantic entities, which also takes account of graph structure for measuring edge importance.

In order to retrieve the most informative knowledge, we first extract the candidate nodes in the ConceptNet by analyzing the prominent visual objects in images with Fast-RCNN BIBREF27 , and textual keywords with the Natural Language Toolkit BIBREF29 . Both of them are then associated with the corresponding semantic entities in ConceptNet BIBREF28 by matching all possible n-grams of words. Afterwards, we retrieve the first-order subgraph using these selected nodes from ConceptNet BIBREF28 , which includes all edges connecting with at least one candidate node. It is assumed that the resultant subgraph contains the most relevant information, which is sufficient to answer questions by reducing the redundancy. The resultant first-order knowledge subgraph is denoted as G .

Finally, we compress the subgraph G by evaluating and ranking the importance of edges in G using a designed score function, and carefully select the top- N edges along with the nodes for subsequent task. Specifically, we first assign initial weights w_{i} for each subgraph node, e.g., the initial weights for visual object can be proportional to their corresponding bounding-box area such that the dominant objects

receive more attention, the textual keywords are treated equally. Then, we calculate the importance score of each node in G by traversing each edge and propagating node weights to their neighbors with a decay factor $r\in (0,1)$ as

$$score(i)=w_{i}+\sum _{j \in G \backslash i} r ^n w_{j},$$ (Eq. 8)

where n is the number of link hops between the entity i and j . For simplicity, we ignore the edge direction and edge type (relation type), and define the importance of edge $w_{i,j}$ as the weights sum of two connected nodes as

$$w_{i,j}=score(i)+score(j), \quad \forall (i,j) \in G.$$ (Eq. 9)

In this paper, we take the top- N edges ranked by $w_{i,j}$ as the final candidate knowledge for the given context, denoted as $G^\ast $.

Knowledge Embedding in Memories

The candidate knowledge that we have extracted is represented in a symbolic triplet format, which is intrinsically incompatible with DNNs. This fact urges us to embed the entities and relation of knowledge triples into a continuous vector space. Moreover, we regard each entity-relation-entity triple as one knowledge unit, since each triple naturally represents one piece of fact. The knowledge units can be stored in memory slots for reading and writing, and distilled through an attention mechanism for the subsequent tasks.

In order to embed the symbolic knowledge triples into memory vector slots, we treat the entities and relations as words, and map them into a continuous vector space using word embedding BIBREF30 .

Afterwards, the embedded knowledge is encoded into a fixed-size vector by feeding it to a recurrent neural network (RNN). Specifically, we initialize the word-embedding matrix with a pre-trained GloVe word-embedding BIBREF30 , and refine it simultaneously with the rest of procedure of question and candidate answer embedding. In this case, the entities and relations share a common embedding space with other textual elements (questions and answers), which makes them much more flexible to fuse later.

Afterwards, the knowledge triples are treated as SVO phrases of $(subject, verb, object)$, and fed to to a standard two-layer stacked LSTM as

$$
\begin{aligned}
&C^{(t)}_{i} = \text{LSTM}\left(\mathbf {L}[w^{t}_{i}], C^{(t-1)}_{i}\right), \\
& t=\lbrace 1,2,3\rbrace , \text{ and } i=1, \cdots , N,\nonumber
\end{aligned}
$$ (Eq. 11)

where w^{t}_{i} is the t_{th} word of the i_{th} SVO phrase, $(w^{1}_{i},w^{2}_{i},w^{3}_{i}) \in G^\ast $, $\mathbf {L}$ is the word embedding matrix BIBREF30 , and C_{i} is the internal state of LSTM cell when forwarding the i_{th} SVO phrase. The rationale lies in the fact that the LSTM can capture the semantic meanings effectively when the knowledge triples are treated as SVO phrases.

For each question-answering context, we take the LSTM internal states of the relevant knowledge triples as memory vectors, yielding the embedded knowledge stored in memory slots as

$$\mathbf {M}=\left[C^{(3)}_{i}\right],$$ (Eq. 12)

where $\mathbf {M}(i)$ is the i_{th} memory slot corresponding to the i_{th} knowledge triples, which can be used for further answer inference. Note that the method is different from the usual knowledge graph embedding models, since our model aims to fuse knowledge with the latent features of

images and text, whereas the alternative models such as TransE BIBREF26 focus on link prediction task.

Attention-based Knowledge Fusion with DNNs

We have stored N relevant knowledge embeddings in memory slots for a given question-answer context, which allows to incorporate massive knowledge when N is large. The external knowledge overwhelms other contextual information in quantity, making it imperative to distill the useful information from the candidate knowledge. The Dynamic Memory Network (DMN) BIBREF22 , BIBREF25 provides a mechanism to address the problem by modeling interactions among multiple data channels. In the DMN module, an episodic memory vector is formed and updated during an iterative attention process, which memorizes the most useful information for question answering. Moreover, the iterative process brings a potential capability of multi-hop reasoning.

This DMN consists of an attention component which generates a contextual vector using the previous memory vector, and an episodic memory updating component which updates itself based on the contextual vector. Specifically, we propose a novel method to generate the query vector $\mathbf {q}$ by feeding visual and textual features to a non-linear fully-connected layer to capture question-answer context information as

$$\mathbf {q} = \tanh \left(\mathbf {W}_{1} \left[\mathbf {f}^{(I)};\mathbf {f}^{(Q)};\mathbf {f}^{(A)}\right]+\mathbf {b}_{1}\right),$$ (Eq. 14)

where $\mathbf {W}_1$ and $\mathbf {b}_{1}$ are the weight matrix and bias vector, respectively; and, $\mathbf {f}^{(I)}$, $\mathbf {f}^{(Q)}$ and $\mathbf {f}^{(A)}$ are denoted as DNN features corresponding to the images, questions and multi-choice answers, respectively. The query vector $\mathbf {q}$ captures information from question-answer context. During the training process, the query

vector $\mathbf{q}$ initializes an episodic memory vector $\mathbf{m}^{(0)}$ as $\mathbf{m}^{(0)}=\mathbf{q}$. A iterative attention process is then triggered, which gradually refines the episodic memory $\mathbf{m}$ until the maximum number of iterations steps $\mathbf{b}_{1}0$ is reached. By the $\mathbf{b}_{1}1$ iteration, the episodic memory $\mathbf{b}_{1}2$ will memorize useful visual and external information to answer the question. Attention component. At the $\mathbf{b}_{1}3$ iteration, we concatenate each knowledge embedding $\mathbf{b}_{1}4$ with last iteration episodic memory $\mathbf{b}_{1}5$ and query vector $\mathbf{b}_{1}6$, then apply the basic soft attention procedure to obtain the $\mathbf{b}_{1}7$ context vector $\mathbf{b}_{1}8$ as

$$\mathbf{z}_{i}^{(t)} &= \left[\mathbf{M}_{i};\mathbf{m}^{(t-1)};\mathbf{q}\right] \\
\alpha^{(t)} &= softmax\left(\mathbf{w}\tanh\left(\mathbf{W}_{2}\mathbf{z}_{i}^{(t)}+\mathbf{b}_{2}\right)\right) \\
\mathbf{c}^{(t)}&=\sum_{i=1}^{N}\alpha_{i}^{(t)}\mathbf{M}_{i} \quad t=1, \cdots, T,$$ (Eq. 15)

where $\mathbf{z}_{i}^{(t)}$ is the concatenated vector for the i_{th} candidate memory at the t_{th} iteration; $\alpha_{i}^{(t)}$ is the i_{th} element of $\alpha^{(t)}$ representing the normalized attention weight for $\mathbf{M}_{i}$ at the t_{th} iteration; and, $\mathbf{w}$, $\mathbf{W}_{2}$ and $i_{\text{th}}0$ are parameters to be optimized in deep neural networks.

Hereby, we obtain the contextual vector $\mathbf{c}^{(t)}$, which captures useful external knowledge for updating episodic memory $\mathbf{m}^{(t-1)}$ and providing the supporting facts to answer the open-domain questions.

Episodic memory updating component. We apply the memory update mechanism BIBREF21, BIBREF25 as

$$\mathbf {m}^{(t)}=ReLU\left(\mathbf {W}_{3} \left[\mathbf {m}^{(t-1)};\mathbf {c}^{(t)};\mathbf {q}\right]+\mathbf {b}_{3}\right),$$ (Eq. 16)

where $\mathbf {W}_{3}$ and $\mathbf {b}_{3}$ are parameters to be optimized. After the iteration, the episodic memory $\mathbf {m}^{(T)}$ memorizes useful knowledge information to answer the open-domain question.

Compared with the DMN+ model implemented in BIBREF25 , we allows the dynamic memory network to incorporate the massive external knowledge into procedure of VQA reasoning. It endows the system with the capability to answer more general visual questions relevant but beyond the image contents, which is more attractive in practical applications.

Fusion with episodic memory and inference. Finally, we embed visual features $\mathbf {f}^{(I)}$ along with the textual features $\mathbf {f}^{(Q)}$ and $\mathbf {f}^{(A)}$ to a common space, and fuse them together using Hadamard product (element-wise multiplication) as

$$\begin{aligned}&\mathbf {e}^{(k)}=\tanh \left(\mathbf {W}^{(k)}\mathbf {f}^{(k)}+\mathbf {b}^{(k)}\right), k \in \lbrace I, Q, A\rbrace \\ &\mathbf {h} =\mathbf {e}^{(I)} \odot \mathbf {e}^{(Q)} \odot \mathbf {e}^{(A)},\end{aligned}$$ (Eq. 17)

 where $\mathbf {e}^{(I)}$, $\mathbf {e}^{(Q)}$ and $\mathbf {e}^{(A)}$ are embedded features for image, question and answer, respectively; $\mathbf {h}$ is the fused feature in this common space; and, $\mathbf {W}^{(I)}$, $\mathbf {W}^{(Q)}$ and $\mathbf {W}^{(A)}$ are corresponding to the parameters in neural networks.

The final episodic memory $\mathbf {m}^{(T)}$ are concatenated with the fused feature $\mathbf {h}$ to

predict the probability of whether the multi-choice candidate answer is correct as

$$ans^* = \operatorname*{arg\,max}_{ans \in \lbrace 1,2,3,4\rbrace } softmax\left(\mathbf {W}_{4}\left[\mathbf {h}_{ans};\mathbf {m}^{(T)}_{ans}\right]+\mathbf {b}_{4}\right),$$ (Eq. 18)

where ans represents the index of multi-choice candidate answers; the supported knowledge triples are stored in $\mathbf {m}^{(T)}_{ans}$; and, $\mathbf {W}_{4}$ and $\mathbf {b}_{4}$ are the parameters to be optimized in the DNNs. The final choice are consequentially obtained once we have $ans^\ast $.

Our training objective is to learn parameters based on a cross-entropy loss function as

$$\mathcal {L} = -\frac{1}{D}\sum _{i}^{D}\big (y_{i}\log \hat{y_{i}}+(1-y_{i})\log (1-\hat{y_{i}})\big),$$ (Eq. 19)

where $\hat{y_{i}}=p_{i}(A^{(i)}|I^{(i)},Q^{(i)},K^{(i)};\theta)$ represents the probability of predicting the answer $A^{(i)}$, given the i_{th} image $I^{(i)}$, question $Q^{(i)}$ and external knowledge $K^{(i)}$; $\theta $ represents the model parameters; D is the number of training samples; and y_{i} is the label for the i_{th} sample. The model can be trained in an end-to-end manner once we have the candidate knowledge triples are retrieved from the original knowledge graph.

Experiments

In this section, we conduct extensive experiments to evaluate performance of our proposed model, and compare it with its variants and the alternative methods. We specifically implement the evaluation on a

public benchmark dataset (Visual7W) BIBREF7 for the close-domain VQA task, and also generate numerous arbitrary question-answers pairs automatically to evaluate the performance on open-domain VQA. In this section, we first briefly review the dataset and the implementation details, and then report the performance of our proposed method comparing with several baseline models on both close-domain and open-domain VQA tasks.

Datasets

We train and evaluate our model on a public available large-scale visual question answering datasets, the Visual7W dataset BIBREF7 , due to the diversity of question types. Besides, since there is no public available open-domain VQA dataset for evaluation now, we automatically build a collection of open-domain visual question-answer pairs to examine the potentiality of our model for answering open-domain visual questions.

The Visual7W dataset BIBREF7 is built based on a subset of images from Visual Genome BIBREF31 , which includes questions in terms of (what, where, when, who, why, which and how) along with the corresponding answers in a multi-choice format. Similar as BIBREF7 , we divide the dataset into training, validation and test subsets, with totally 327,939 question-answer pairs on 47,300 images. Compared with the alternative dataset, Visual7W has a diverse type of question-answer and image content BIBREF13 , which provides more opportunities to assess the human-level capability of a system on the open-domain VQA.

In this paper, we automatically generate numerous question-answer pairs by considering the image content and relevant background knowledge, which provides a test bed for the evaluation of a more realistic VQA task. Specifically, we generate a collection automatically based on the test image in the Visual7W by filling a set of question-answer templates, which means that the information is not present

during the training stage. To make the task more challenging, we selectively sample the question-answer pairs that need to reasoning on both visual concept in the image and the external knowledge, making it resemble the scenario of the open-domain visual question answering. In this paper, we generate 16,850 open-domain question-answer pairs on images in Visual7W test split. More details on the QA generation and relevant information can be found in the supplementary material.

Implementation Details

In our experiments, we fix the joint-embedding common space dimension as 1024, word-embedding dimension as 300 and the dimension of LSTM internal states as 512. We use a pre-trained ResNet-101 BIBREF32 model to extract image feature, and select 20 candidate knowledge triples for each QA pair through the experiments. Empirical study demonstrates it is sufficient in our task although more knowledge triples are also allowed. The iteration number of a dynamic memory network update is set to 2, and the dimension of episodic memory is set to 2048, which is equal to the dimension of memory slots.

In this paper, we combine the candidate Question-Answer pair to generate a hypothesis, and formulate the multi-choice VQA problem as a classification task. The correct answer can be determined by choosing the one with the largest probability. In each iteration, we randomly sample a batch of 500 QA pairs, and apply stochastic gradient descent algorithm with a base learning rate of 0.0001 to tune the model parameters. The candidate knowledge is first retrieved, and other modules are trained in an end-to-end manner.

In order to analyze the contributions of each component in our knowledge-enhanced, memory-based model, we ablate our full model as follows:

KDMN-NoKG: baseline version of our model. No external knowledge involved in this model. Other

parameters are set the same as full model.

KDMN-NoMem: a version without memory network. External knowledge triples are used by one-pass soft attention.

KDMN: our full model. External knowledge triples are incorporated in Dynamic Memory Network.

We also compare our method with several alternative VQA methods including (1) LSTM-Att BIBREF7 , a LSTM model with spatial attention; (2) MemAUG BIBREF33 : a memory-augmented model for VQA; (3) MCB+Att BIBREF6 : a model combining multi-modal features by Multimodal Compact Bilinear pooling; (4) MLAN BIBREF11 : an advanced multi-level attention model.

Results and Analysis

In this section, we report the quantitative evaluation along with representative samples of our method, compared with our ablative models and the state-of-the-art method for both the conventional (close-domain) VQA task and open-domain VQA.

In this section, we report the quantitative accuracy in Table 1 along with the sample results in 3 . The overall results demonstrate that our algorithm obtains different boosts compared with the competitors on various kinds of questions, e.g., significant improvements on the questions of Who (5.9%), and What (4.9%) questions, and slightly boost on the questions of When (1.4%) and How (2.0%). After inspecting the success and failure cases, we found that the Who and What questions have larger diversity in questions and multi-choice answers compared to other types, therefore benefit more from external background knowledge. Note that compared with the method of MemAUG BIBREF33 in which a memory mechanism is also adopted, our algorithm still gain significant improvement, which further

confirms our belief that the background knowledge provides critical supports.

We further make comprehensive comparisons among our ablative models. To make it fair, all the experiments are implemented on the same basic network structure and share the same hyper-parameters. In general, our KDMN model on average gains 1.6% over the KDMN-NoMem model and 4.0% over the KDMN-NoKG model, which further implies the effectiveness of dynamic memory networks in exploiting external knowledge. Through iterative attention processes, the episodic memory vector captures background knowledge distilled from external knowledge embeddings. The KDMN-NoMem model gains 2.4% over the KDMN-NoKG model, which implies that the incorporated external knowledge brings additional advantage, and act as a supplementary information for predicting the final answer. The indicative examples in Fig. 3 also demonstrate the impact of external knowledge, such as the 4th example of "Why is the light red?". It would be helpful if we could retrieve the function of the traffic lights from the external knowledge effectively.

In this section, we report the quantitative performance of open-domain VQA in Table 2 along with the sample results in Fig. 4 . Since most of the alternative methods do not provide the results in the open-domain scenario, we make comprehensive comparison with our ablative models. As expected, we observe that a significant improvement (12.7%) of our full KDMN model over the KDMN-NoKG model, where 6.8% attributes to the involvement of external knowledge and 5.9% attributes to the usage of memory network. Examples in Fig. 4 further provide some intuitive understanding of our algorithm. It is difficult or even impossible for a system to answer the open domain question when comprehensive reasoning beyond image content is required, e.g., the background knowledge for prices of stuff is essential for a machine when inferring the expensive ones. The larger performance improvement on open-domain dataset supports our belief that background knowledge is essential to answer general visual questions. Note that the performance can be further improved if the technique of ensemble is allowed. We fused the results of several KDMN models which are trained from different initializations.

Experiments demonstrate that we can further obtain an improvement about 3.1% .

Conclusion

In this paper, we proposed a novel framework named knowledge-incorporate dynamic memory network (KDMN) to answer open-domain visual questions by harnessing massive external knowledge in dynamic memory network. Context-relevant external knowledge triples are retrieved and embedded into memory slots, then distilled through a dynamic memory network to jointly inference final answer with visual features. The proposed pipeline not only maintains the superiority of DNN-based methods, but also acquires the ability to exploit external knowledge for answering open-domain visual questions. Extensive experiments demonstrate that our method achieves competitive results on public large-scale dataset, and gain huge improvement on our generated open-domain dataset.

Details of our Open-domain Dataset Generation

We obey several principles when building the open-domain VQA dataset for evaluation: (1) The question-answer pairs should be generated automatically; (2) Both of visual information and external knowledge should be required when answering these generated open-domain visual questions; (3) The dataset should in multi-choices setting, in accordance with the Visual7W dataset for fair comparison.

The open-domain question-answer pairs are generated based on a subset of images in Visual7W BIBREF7 standard test split, so that the test images are not present during the training stage. For one particular image that we need to generate open-domain question-answer pairs about, we firstly extract several prominent visual objects and randomly select one visual object. After linked to a semantic entity in ConceptNet BIBREF28 , the visual object connects other entities in ConceptNet through various relations, e.g. UsedFor, CapableOf, and forms amount of knowledge triples $(head, relation, tail)$, where either

$head$ or $tail$ is the visual object. Again, we randomly select one knowledge triple, and fill into a $relation$ -related question-answer template to obtain the question-answer pair. These templates assume that the correct answer satisfies knowledge requirement as well as appear in the image, as shown in table 3 .

For each open-domain question-answer pair, we generate three additional confusing items as candidate answers. These candidate answers are randomly sampled from a collection of answers, which is composed of answers from other question-answer pairs belonging to the same $relation$ type. In order to make the open-domain dataset more challenging, we selectively sample confusing answers, which either satisfy knowledge requirement or appear in the image, but not satisfy both of them as the ground-truth answers do. Specifically, one of the confusing answers satisfies knowledge requirement but not appears in image, so that the model must attend to visual objects in image; another one of the confusing answers appears in the image but not satisfies knowledge requirement, so that the model must reason on external knowledge to answer these open-domain questions. Please see examples in Figure 5 .

In total, we generate 16,850 open-domain question-answer pairs based on 8,425 images in Visual7W test split.