Textbook Question Answering under Instructor Guidance with Memory Networks

Juzheng Li[1] Hang Su[1] Jun Zhu[1] Siyu Wang[2] Bo Zhang[1]
[1] Department of Computer Science and Technology, Tsinghua Lab of Brain and Intelligence
[1] Beijing National Research Center for Information Science and Technology, BNRist Lab
[1,2] Tsinghua University, 100084 China
lijuzheng09@gmail.com; {suhangss,dcszj,dcszb}@tsinghua.edu.cn; siyuwang19@mails.tsinghua.edu.cn

Abstract

Textbook Question Answering (TQA) is a task to choose the most proper answers by reading a multi-modal context of abundant essays and images. TQA serves as a favorable test bed for visual and textual reasoning. However, most of the current methods are incapable of reasoning over the long contexts and images. To address this issue, we propose a novel approach of Instructor Guidance with Memory Networks (IGMN) which conducts the TQA task by finding contradictions between the candidate answers and their corresponding context. We build the Contradiction Entity-Relationship Graph (CERG) to extend the passage-level multi-modal contradictions to an essay level. The machine thus performs as an instructor to extract the essay-level contradictions as the Guidance. Afterwards, we exploit the memory networks to capture the information in the Guidance, and use the attention mechanisms to jointly reason over the global features of the multi-modal input. Extensive experiments demonstrate that our method outperforms the state-of-the-arts on the TQA dataset. The source code is available at https://github.com/freerailway/igmn.

1. Introduction

Question Answering (QA) has been a significant research branch in natural language processing (NLP) in the past decades [13], which aims to answer questions given a specific textual narrative. Thanks to the availability of large-scale QA datasets with additional visual supporting information [23, 34, 2], Visual Question Answering (VQA) has attracted a substantial interest from the community of computer vision [10, 30, 8]. Recently, a new task of Textbook Question Answering (TQA) is proposed which aims to answer arbitrary questions by reading a large context [16]. Different from the machine comprehension with languages or visual question answering with images, TQA consists of abundant essays and images and it requires to learn and comprehend multi-modal sources of information across text documents and images, as illustrated in Figure 1. TQA therefore connects computer vision and natural language processing and pushes forward boundaries of both fields.

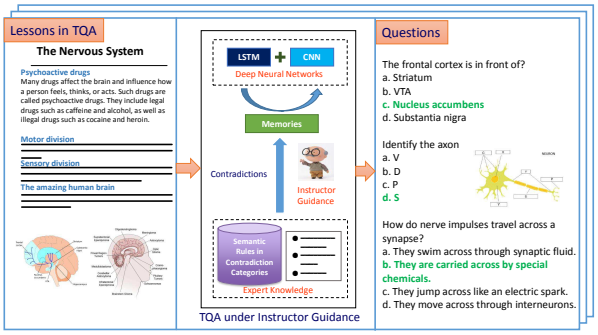


Figure 1: An example for the TQA task. Given a multi-model context (*e.g.*, abundant essays and images), the instructor first extracts the local features, following categories of contradictions to provide the Guidance. Then the attention-based memory networks jointly reason from the global features and the Guidance to obtain the most proper choice for each question.

The TQA task brings new challenges. As the most distinctive characteristic, TQA uses long essays to describe concepts. For example, the essays in TQA have an average length of 1800 words. Most of the previous models for textual machine comprehension (e.g., bidirectional attention flow mechanism [24], the multi-layer embedding method with memory networks [21] and the R-Net [28]) can cope with passages with about 300 words. The existing methods are not eligible for such long essays even if additional attention, selection or truncation mechanisms are applied, because the facts to answer a question are often dispersed in the essay in the TQA task [16].

We notice that recent machine comprehension systems witness a progress on syntax and semantic analysis for abstracting over the word order differences [27, 4, 31]. How-

ever, such local features are difficult to be summarized over the whole essay, especially incorporated with the image information. We then find a unique type of semantic relations: the contradiction, to conduct this problem. A contradiction describes that two expressions can not be true at the same time. Contradictions are easy to be summarized over long essays, considering that opposite expressions are still conflicting after affiliating some reconcilable statements. Nevertheless, the capabilities of extracting language information by semantic-based algorithms are limited by the scales of pre-defined semantic rules [31]. It is also difficult to incorporate the image features into the semantic-based algorithms since the structural and grammatical rules cannot describe every details of images.

The memory networks provide a good opportunity to deal with the global information of the large context since it allows a model to implement reasoning within a possibly large external memory. Memory networks extend the scale of data that the network can read in each time step, and consequently obtain promising performance in both textual machine comprehension [12, 29, 32] and vision-related tasks such as image captioning [22] and scene labeling [1]. It inspires us to adopt the memory mechanism to cope with long essay in the TQA tasks, which also provides an advantage to reason over the contradictions. We name the contradictions found by the instructor as the Guidance and write it in the memory for further reasoning.

1.1. Our Proposal

We propose the Instructor Guidance with Memory Networks (IGMN), a novel framework that utilizes semantic analysis to comprehend large contexts and suggest the Guidance for the deep neural networks to reason from multi-modal data.

To solve the task in the perspective of finding contradictions, we first change the questions into statements by filling the blanks with possible answers. As the TQA dataset provides candidate answers, this step becomes quite easy. We then find contradictions between the statements and their multi-modal contexts. Recent methods of finding contradictions are usually designed for short paragraphs [17, 20] but are difficult to be extended to deal with the images or long essays because the key points may be distributed in different areas. To address this issue, we propose a new discrete structure of Contradiction Entity-Relationship Graph (CERG) to represent the facts in the context that may lead to contradictions. The structure of the CERG is specially designed for finding contradictions. Based on Marneffe *et al*. [7], we propose new categories of contradictions that underline the fusion of small facts among different parts of the essays and images. We summarize the contradictions that need facts from different areas of the context into two types, Causality and Structure. The links in the CERG model the dispersive facts behind the two types of contradictions. For the contradictions that only need local facts, we summarize them into two types of Entity Distinction and Negation. The facts behind the local contradictions are represented by the nodes of the CERG.

To build the CERGs, we first use the Stanford Parser [26, 19] to acquire dependency parse trees as syntax analysis. Subsequently we retrieve useful entities and their relationships with pre-defined rules specific to each category of contradictions. In this step we use the Natural Language Toolkit (NLTK) [5] for word tokenization, sentence segmentation and word stemming, and use the Stanford CorefAnnotator [6] to annotate the co-references. We then organize the entities in the CERGs with their relationships properly formalized. The essential facts in the long essays are therefore represented in a structured form. As the CERGs are built according to the contradiction categories, the final contradictions are consequently derived by comparing two CERGs. But the derived contradictions are incomplete because such discrete structures cannot capture every details of the large context. Thereupon, we regard the CERG as an "instructor" who can only provide the imperfect contradictions, named as the "Guidance".

Moreover, we use an attention-based memory network to jointly reason the latent facts from the multi-modal context and the Guidance. We record the Guidance in the memory, which is different from the previous methods [16, 21] that exploit input data or instructional rules as repository. With the help of the Guidance to attend to the areas of the context where may exist contradictions, the deep neural networks in our method obtain more direct instructions. The whole method thus captures both discrete and global features of the given context and performs a better reasoning. Our method has an outstanding performance in both of the text-only sub-task and the image engaged task of TQA comparing to various state-of-the-art methods.

To sum up, our main contributions are as follows:

- We propose a novel framework for the textbook question answering (TQA) task by addressing the two challenges of large context and multi-modal data;
- We propose the CERG and new categories of contradictions that enable us to summarize the large textual context as well as the visual concepts;
- We exploit the memory and attention mechanisms to combine the semantic-based and DNN-based methods for deep reasoning.

2. Methodology

In this section we first give an overview of IGMN, and then introduce the details of the two functional modules

including the Instructor-Guided Knowledge Extraction and Answer Generation by Joint Reasoning.

2.1. Overview

Figure 2 illustrates how our method addresses the TQA task. The goal is to choose the most proper answer corresponding to each question with a given multi-modal context. We propose a hierarchical model for this task: the answer is selected by an attention-based memory network in the module of Answer Generation by Joint Reasoning (AGJR, the upper half of Figure 2). AGJR integrates the Guidance obtained from the module of Instructor-Guided Knowledge Extraction (IGKE, the lower half of Figure 2).

In order to extract the useful information from the large context in TQA, we use the IGKE module which provides a Guidance by semantic analysis. Specifically, we focus on the contradiction, a unique type of semantic relations. We use the Contradiction Entity-Relationship Graph (CERG) which facilitates the comparison between large contexts. The Causality and Structure contradictions are the main focus of this paper and we use the CERG to find them. The passage-level contradictions are not our major concern since they are being fully-addressed by existing methods [17, 20, 28, 2], but we still provide an implementation using discrete semantic and visual analysis to show the practicability of our framework. We finally match the CERGs of the question and its corresponding context to obtain the incomplete contradictions, *i.e*. the Guidance.

The IGKE module performs as an instructor to provide the Guidance for the AGJR module. The AGJR module aims at extracting and combining global features from multi-modal inputs to reason out the proper answer. To this end, we use the deep neural networks aided by the memory and attention mechanisms [3]. In particular, BiLSTM [11] and VGG Net [25] are employed to extract the latent facts from the textual and visual data respectively. Then we specially exploit the memory to store the Guidance from the IGKE module, and then use the attention mechanism to reason from the diverse latent facts. Finally the results produced by the attention mechanism of all the options for a question are orderly fed into a LSTM [14] decoder to obtain the final choice.

The two modules in our method are functionally complementary. The IGKE module represents the discrete facts with the CERG, supplementing to the AGJR module that is weak in handling the large context. The AGJR module extracts the global features, which captures more detailed information. We use the memory networks to fuse the two modules, which is also an important progress for the TQA task.

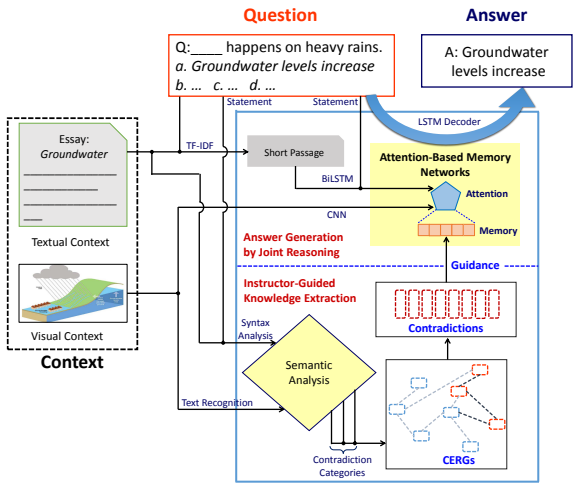


Figure 2: Overall architecture of our proposed method, the Instructor Guidance with Memory Networks (IGMN). The lower part of the figure is the module of Instructor-Guided Knowledge Extraction (IGKE), which represents facts in the long essays and images with the Contradiction Entity-Relationship Graphs (CERGs). The upper part is the module of Answer Generation by Joint Reasoning (AGJR), which accesses the Guidance under a memory network and consequently generates answers by reasoning over the integrated latent facts accordingly by the attention mechanisms.

2.2. Instructor-Guided Knowledge Extraction

In this section we introduce the method to obtain the Guidance. To address the long-essay issue, we propose contradiction categories and embed the facts into CERGs with semantic rules following the categories. We also build CERGs with visual context using spatial analysis rules. Finally we match the CERGs of the questions and their corresponding contexts to find the incomplete contradictions as the Guidance for sequential reasoning by the deep neural networks.

2.2.1 Contradiction Categories

Marneffe *et al*. [7] have summarized different categories of contradictions in text, but some of them are not very applicable for long essays and images. In order to deal with each type more conveniently, we propose 5 new categories based on their work. Table 1 shows the categories and example sentences selected from the TQA dataset.

Firstly, we combine all the word level contradictions including (1), (3) and (6) in Table 1 into the *Entity Distinction* category, which means the hypothesis putting the wrong entity in the certain position in the sentences. Secondly, we change (4) into *Causality* and reserve (5) as *Structure*. *Causality* consists of contradictions in temporal relationships such that one is the reason for another and *Structure* the static relationships such that one is on the left of another. Thirdly, we reserve (2) *Negation* and (7) *World Knowledge*

ID	Type (Marneffe *et al*. [7])	Type (Ours)	Text	Hypothesis
1	Antonym	Entity Distinction	A nation with a lot of neodymium may export that resource to other countries that will import it.	A nation with a lot of neodymium may import that resource to other countries that will export it.
2	Negation	Negation	Soils can also be contaminated if too much salt accumulates in the soil or where pollutants sink into the ground.	Soil will not damage soils.
3	Numeric	Entity Distinction	According to the Big Bang theory, the universe began about 13.7 billion years ago.	According to the Big Bang theory, the universe began 3.7 billion years ago.
4	Factive	Causality	These features are well displayed in the East African Rift, where rifting has begun, and in the Red Sea, where water is filling up the basin created by seafloor spreading.	The Red Sea basin was created by subduction.
5	Structure	Structure	Ships at sea empty their wastes directly into the ocean, for example.	Ships at sea treat their wastes and bring their trash back to land for recycling.
6	Lexical	Entity Distinction	Water is attracted to the soil particles, and capillary action, which describes how water moves through porous media, moves water from wet soil to dry areas.	Water moves through pores from wet soil to dry areas by solvency.
7	World Knowledge	World Knowledge	If the orbital period of a planet is known, then it is possible to determine the planets distance from the Sun.	The planet in the solar system with the largest orbital period is Jupiter.

Table 1: Examples of contradiction types. (1) *Export* and *import* is a pair of antonym. (2) The negation word *not* is used. (3) The numbers are different. (4) The two reasons are different. (5) The two places are different. (6) *Solvency* has not the same meaning as *capillary action*. (7) The essay states that the planet with the largest orbital period is farthest from the sun, but does not tell which planet is the farthest.

ID	Type	Text	Entity
1	Noun Phrase	This is the magnetic evidence.	magnetic.evidence
2	Possessive	Its altitude was 176 km (109 miles) above Earth's surface.	Earth.surface
3	Prepositional Phrase	Slump may be caused by a layer of slippery.	slippery.layer
4	Action	Africa collided with Eurasia to create the Alps.	Africa.collided
5	Compound	Eleven reactors were automatically shut down.	down.shut

Table 2: Types of situations that need to recognize entities. (1-3) The headword is grammatically defined or modified; (4) The action has a subject; (5) Two words are compounded to express a single meaning.

the same name. *Negation* means a explicit negative word is used in the sentence. *World Knowledge* means the contradiction cannot be found only with given sentences and requires external materials.

We now illustrate the intention of our new categories. We conclude the transmissible contradictions as Causality, with similar concept as "deriving (denote as $\rightarrow$)", and Structure, with similar concept as "belonging (denote as $\in$)". The facts behind the two types of contradictions are easily transmitted or combined, such as:

$$\begin{aligned} \mathrm{A} \rightarrow \mathrm{B}\ \mathit{And}\ \mathrm{B} \rightarrow \mathrm{C} \Rightarrow \mathrm{A} \rightarrow \mathrm{C}, \\ \mathrm{A} \in \mathrm{B}\ \mathit{And}\ \mathrm{B} \in \mathrm{C} \Rightarrow \mathrm{A} \in \mathrm{C}, \\ \mathrm{A} \rightarrow \mathrm{B}\ \mathit{And}\ \mathrm{C} \in \mathrm{B} \Rightarrow \mathrm{A} \rightarrow \mathrm{C}, \end{aligned} \tag{1}$$

where the first two chains denote the transitivity of the two types of facts respectively and the third chain combines the two types which means if A results in B, then A results in every part of B. We underline that by observation, the "deriving" and "belonging' relationships takes an significant place in the underlying logics of the nature languages, although they seems so trivial from the view of the logic system. Besides, we conclude the local contradictions as Negation and Entity Distinction. Negation is specially proposed because it is usually disposed independently. The two types of contradictions are not the emphasis of our framework because they can be addressed in the passage level.

2.2.2 Contradiction Entity-Relationship Graph

We now introduce the Contradiction Entity-Relationship Graph (CERG). CERG models the cognitive concepts and their relations described by the large context. CERG uses a discrete structure to embody the facts behind the contradictions. That is to say, CERG records or transmits facts rather than contradictions themselves, but comparing CERGs will arrive at contradictions as CERGs are built following their types.

A node of the CERG denotes an exclusive concept that is represented by the joint meaning of a sequence of words. This structure is corresponding to the Entity Distinction type of contradiction. We ignore the polysemy for simplicity and consider the corpus $\mathcal{X}$ as the set of the words with different meanings. Then a node of the CERG is an ordered sequence $\{a_i\}_{i=1}^{n}$, $a_i \in \mathcal{X}$ with an arbitrary length n. Practically, we distinguish the words in the given context to make every node have a distinctive meaning. We summarize the 5 types as introduced in Table 2. All the 5 types can be recognized in a passage level. Notice that even an action is regarded as an entity, because we aim to model all the linguistic concepts other than Structure and Causality.

A link of the CERG denotes either a "deriving" or "belonging" concept. The concepts are recognized in a similar way of recognizing an entity, also in a passage level. These concepts intuitively models the two fundamental relationships of natural languages: the temporal and the static relationships, corresponding to the contradictions *Causality* and *Structure* introduced in Table 1. The temporal relationship, notated as $\mathrm{A} \rightarrow \mathrm{B}$, indicates that after the happening or appearing of A we are likely to observe the happening

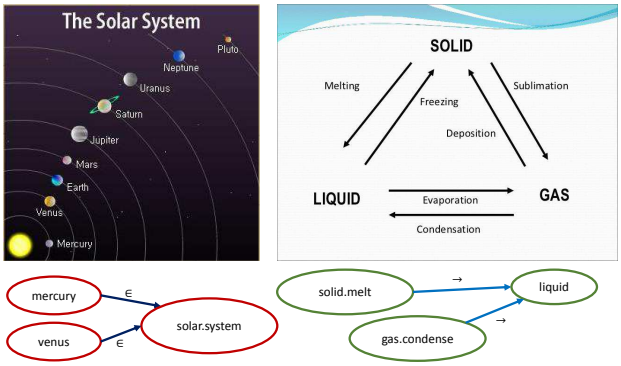


Figure 3: An Example of building the two relationships of the CERG from an image. The left part shows "belonging". We use the positions of texts (here the header position) to speculate their relationships. The right part shows "deriving". We check the relative positions of texts and arrows to speculate their relationships.

or appearing of B. We briefly call it "deriving". The static relationship, notated as A $\in$ B, means that A is part of B, either spatially or abstractly. We briefly call it "belonging". Although recognized locally, the two relationships can link all the entities throughout the essays, which is the key point of extending passage-level contradictions to an essay level.

After we recognize the relationships, we mark the negations on the nodes and links. Then we aggregate all the nodes and links from the whole TQA context as a CERG. For every option in the questions, we fill in the blanks and build a CERG of the statement, querying the whole context CERG to obtain the contradictions.

The CERG is also proper for images, especially for those diagrams with isolated visual concepts. The consistency also results from the fundamentality of the temporal and static relationships. Images usually present relationships with simple logics. The Structure type of contradictions can describe most of the situations in an image that are different from its description, such as the positions and appearances. For diagram-like images, the Causality contradictions are enough to describe the errors in the presented evolutions. Practically, we detect visual concepts as entities and speculate their possible relationships by their relative and absolute positions, whether linked by arrows, etc. Figure 3 shows an example. Presently we only detect texts in the images using existing methods such as TextBoxes [18]. But our framework is capable for further use if stick figures can also be properly recognized.

2.3. Answer Generation by Joint Reasoning

This module aims at combining multi-modal data from the question, the textual and visual context and the Guidance to reason out the most proper choice. We exploit memory to store the Guidance. We use BiLSTM and CNN respectively to extract essential facts from the input, and use the attention mechanisms to merge the latent features for reasoning. The reasoning result is fed into an LSTM decoder to obtain the final answer.

2.3.1 Construction of Guidance Memory

We exploit the memory as repository of the Guidance from the IGKE module to capture the long-term information (Figure 4). We use $\mathbf{u}^G = \{u_i^G\}_{i=1}^D$ to denote the Guidance represented in the memory in size D where each element is a word embedding of dimension 100. The memory vectors can be calculated by

$$\begin{aligned}
\mathbf{m}_j^a &= \mathrm{ReLU}(\mathbf{W}_h^a \mathbf{u}_j^G + \mathbf{b}_h^a), \\
\mathbf{m}_j^c &= \mathrm{ReLU}(\mathbf{W}_h^c \mathbf{u}_j^G + \mathbf{b}_h^c), \\
\mathbf{M}^a &= [\mathbf{m}_1^a; \mathbf{m}_2^a; ...; \mathbf{m}_D^a], \\
\mathbf{M}^c &= [\mathbf{m}_1^c; \mathbf{m}_2^c; ...; \mathbf{m}_D^c],
\end{aligned} \tag{2}$$

where superscripts a and c denote the input (*i.e.* addressing) and output of the memory respectively following [29, 22]; $\mathbf{m}^{a/c}$ are the input/output memory representations at every slot of dimension 512; $\mathbf{M}^{a/c} \in \mathbb{R}^{D\times 512}$ are the total input/output memory representations; $\mathbf{W}_h^{a/c} \in \mathbb{R}^{512\times 100}$ and $\mathbf{b}_h^{a/c}$ are trainable parameters; the operator $[;]$ means direct concatenation; $\mathrm{ReLU}()$ means the ReLU [9] activation function. Eq. (2) makes the raw Guidance capable for reading, which serves as a support for the following question answering.

Let $\mathbf{e}^Q = \{e_t^Q\}_{t=1}^n$ denote the word embeddings of the statement sentence with the length n. When consulting the encoded features from the statement, we obtain the final output by

$$\begin{aligned}
\mathbf{u}_t^Q &= \mathrm{BiLSTM}_Q(\mathbf{u}_{t-1}^Q, \mathbf{e}_t^Q), \\
\mathbf{q}_t &= \mathrm{ReLU}(\mathbf{W}_q \mathbf{u}_t^Q + \mathbf{b}_q), \\
\mathbf{p}_t &= \mathrm{softmax}(\mathbf{M}_t^a \mathbf{q}_t), \\
\mathbf{M}_t^o &= \mathbf{p}_t \circ \mathbf{M}^c,
\end{aligned} \tag{3}$$

where $\mathbf{u}^Q \in \mathbb{R}^{n\times 256}$ denotes the encoded features with dimension 256 from the statement $\mathbf{e}^Q$; $\mathbf{W}_q \in \mathbb{R}^{512\times 256}$ and $\mathbf{b}_q$ are trainable parameters; the operator $\circ$ means element-wise multiplication; $\mathbf{q}_t \in \mathbb{R}^{512}$ is the query vector at time t to match the memory, and $\mathbf{p}_t \in \mathbb{R}^D$ is an attention mask to select the most proper part of the memory to attend at time t; $\mathrm{BiLSTM}()$ means the BiLSTM [11] encoder. We finally obtain $\mathbf{M}^o \in \mathbb{R}^{n\times 512}$ as the reasoned results compared between the facts in the statement and the Guidance.

2.3.2 Attention-Based Joint Reasoning

Then we use the soft attention mechanisms to further merge the extracted facts of the image, passage with the Guidance

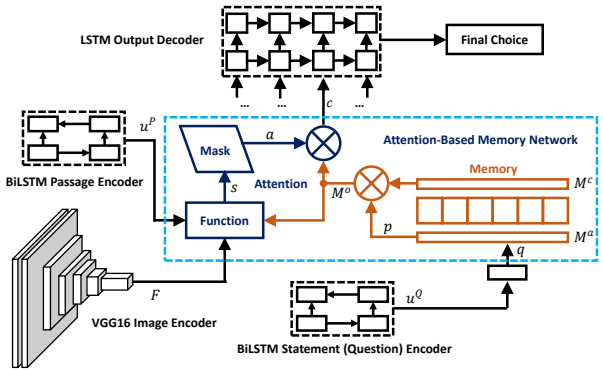


Figure 4: The attention-based memory network in our model. The encoded statement is made as a query to visit the memory to obtain the memory output. The attention mechanism combines the encoded passage and image to generate a mask acting on the memory output. The latent features of all the options for a question are aggregated orderly to find the final choice by an LSTM decoder.

output generated by Eq. (3). The structure of the mechanisms is shown in Figure 4.

Let $\mathbf{e}^P = \{e_t^P\}_{t=1}^m$ denote the word embeddings of the short passage with a total length of m, we encode the short passage by

$$\mathbf{u}_t^P = \mathrm{BiLSTM}_P(\mathbf{u}_{t-1}^P, \mathbf{e}_t^P). \tag{4}$$

We use the VGG16 [25] network to extract features from the images. Let $\mathbf{F} \in \mathbb{R}^{7\times7\times512}$ be the pool5 feature maps of VGG16, $\mathbf{M}^o$ be the output of the memory queried by $\mathbf{u}^Q$ following Eq. (3). Our attention mechanisms can be expressed as:

$$\begin{aligned} \mathbf{s}_j^t &= \mathbf{v}^{\mathrm{T}}\tanh(\mathbf{W}^Q\mathbf{M}_j^o + \mathbf{W}^P\mathbf{u}_t^P + \mathbf{W}^F\mathbf{F}), \\ \mathbf{a}^t &= \mathrm{softmax}(\mathbf{s}^t), \\ \mathbf{c}^t &= \mathbf{a}^t \circ \mathbf{M}^o, \end{aligned} \tag{5}$$

where $\mathbf{v}$, $\mathbf{W}^Q$, $\mathbf{W}^P$ and $\mathbf{W}^F$ are trainable weights. Hereby, the support Guidance is stored in $\mathbf{M}^o$; $\mathbf{s}$ is the result of the attention function to attend the most proper part of the memory output and $\mathbf{a}$ is the normalized attention mask. $\mathbf{c} = \{c^t\}_{t=1}^n$ performs as the final reasoning result for a certain option to be fed into an LSTM decoder orderly. Thus, following the steps in the Eq. (2)∼(5), the Guidance is first integrated with the statement, and then the passage along with the image content, deriving the final suggested choice (Figure 4).

3. Experiments

We now present the experiments to validate the effectiveness of IGMN. We first introduce the dataset and the existing and ablation methods that we will compare with. Next, by comparing the quantitative results, we show the advantages of our overall framework and various modules. Finally we present examples that qualitatively demonstrate the ability of IGMN.

3.1. Datasets

We use the TQA dataset [16], which consists of 1,076 lessons downloaded from middle school on-line curricula[1] split to training, validation and test sets. The training set has 15,154 questions and 666 lessons with an average length of about 1,700 words. The validation set has 5,309 question and 200 lessons with an average length of about 1,900 words. Because the answers of the test set is hidden, we use the validation set to evaluate all the methods.

3.2. Compared Methods and Results

We compare our method with several alternative methods for QA or VQA task including:

MemN+VQA This method is proposed as a baseline method for TQA task [16]. It uses LSTM with a mechanism of Memory Networks [29] to process texts in essays and questions. For diagram questions, it employs popular VQA mechanisms like [33], which encodes images with a VGG network [25] and then inserts the encoded features into the memory.

MemN+VQA+HT This method is similar to MemN+VQA. The only difference is that the relative passages are not properly selected. Instead, the first sentences and the last sentence of the relative essay constitute each passage for the corresponding question.

MemN+DPG This method is also proposed by [16]. It uses a mechanism DPG based on DSDP-NET [15] which translates images into nature language sentences. DPG is short for Diagram Parse Graph, which models the structure of images as proposed in [15]. Then the method combines textual and visual sentences with Memory Networks.

BiDAF+DPG This end-to-end method uses a bi-directional attention mechanism [24] to capture dependencies between question and corresponding context paragraph [16]. It uses DPG to translate diagrams into sentences and combines them with essays.

Challenge The best results in a recent competition on TQA task[2]. As the organizer reports, Haurilet and Al-Halah won the text-question track and Tay and Luu won the diagram-question track. We merge their best results.

Random Randomly choosing an option for all the questions.

In order to analyze the contributions of each component of our IGMN framework, we ablate our full model as follows:

IGMN-AGJR A method for ablation analysis. This is our method without Instructor-Guided Knowledge Extracting module. Deep neural networks process textual and vi-

[1] http://www.ck12.org
[2] http://vuchallenge.org/tqa.html

Model	Text T/F	Text MC	Text All	Diagram	All
Random	50.10	22.88	33.62	24.96	29.08
MemN+VQA [16]	50.50	31.05	38.73	31.82	35.11
MemN+VQA+HT	50.30	28.10	36.87	29.81	33.17
MemN+DPG [16]	50.50	30.98	38.69	32.83	35.62
BiDAF+DPG [16]	50.40	30.46	38.33	32.72	35.39
Challenge	-	-	45.57	35.85	40.48
IGMN-AGJR	51.20	30.26	38.53	31.25	34.71
IGMN-EntD	51.60	32.55	40.07	31.75	35.71
IGMN-IGKE	55.31	38.76	45.29	34.77	39.78
IGMN (ours)	**57.41**	**40.00**	**46.88**	**36.35**	**41.36**

Table 3: Detailed results (% accuracy) of different types of questions in the TQA task are presented. We present the accuracies of true or false text-only questions (Text T/F), multiple choice text-only questions (Text MC), all the text-only questions (Text All = Text T/F ∪ Text MC), multiple choice questions with images (Diagram) and total questions (All = Text All ∪ Diagram).

sual data without referring to the memory containing the Guidance.

IGMN-EntD A method for ablation analysis. We do not build the CERGs and only compare the distinct entities between the question and the relative context.

IGMN-IGKE A method for ablation analysis. We simply regard the Guidance as the final answers without the help of deep neural networks.

IGMN The entire framework introduced by this paper.

3.3. Overall Results on TQA

The overall results on TQA dataset are shown in Table 3. We observe that our method has outstanding results in accuracy in every type of questions.

Although IGMN attains more accuracy improvement in Text MC (for about 9% than baselines) than in Text T/F (for about 7% than baselines), we believe the improvement in Text T/F is more significant considering that the baselines have similar results to random in Text T/F. The results suggest the stronger capability of our method on judgment by finding contradictions.

We can also observe that IGMN attains considerable progress compared to MemN+VQA and MemN+DPG, which also employ memory networks but as repository for original context. This comparison shows the significant effect of our Instructor-Guided Knowledge Extraction module to obtain the contradictions by matching the CERGs from the given context. The information embedded in the long essays is explored effectively by our proposed CERGs, which further provides supports for the subsequential question answering.

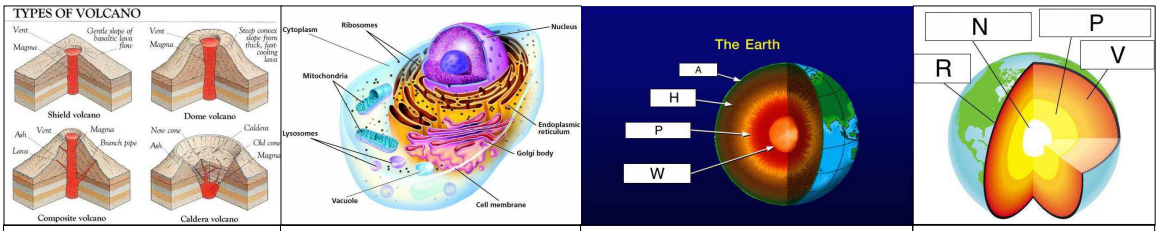
The results show that the improvement made in diagram questions by IGMN is not so significant as in text questions. This is possibly because we only recognize text in the images in the IGKE module and utilize the graphic information in the AGJR module. But the accuracies of IGMN in Diagram questions still surpasses VQA and DPG, suggesting that CERGs are so powerful that the model only considering partial situations is also able to perform well.

IGMN also surpasses the newly released challenge results in both text-only and image engaged questions, although the methods of the challenge are not published so the comparison is only for reference. The challenge allows ensemble but IGMN is a single model. If we use ensemble, we obtain 2.2% accuracy improvement than the challenge results of all the questions, which is again significant.

3.4. Ablation Analysis

The results of ablation experiments in Table 3 demonstrate that the two modules in our method work synergistically. We can observe apparent decline of the accuracy when any of the two modules is ablated. Moreover, it is not surprising that the effect of IGMN mainly comes from the Guidance. Without the Guidance, the AGJR module cannot provide better results than the baselines, in contrast that the IGKE module itself attains about 4% improvement. But the AGJR module plays an important supplementary role. Comparing IGMN with IGMN-IGKE there is 1.58% improvement of the overall accuracy to combine with the attention-based memory networks than the Guidance only. This suggests that the method to obtain the Guidance contains much omission and the attention-based memory networks can fill up some of the details by learning from abundant training data. Comparing IGMN with IGMN-AGJR, there is 6.65% improvement of the overall accuracy which suggests that IGMN is able to utilize the facts in the Guidance effectively with memory networks.

Comparing IGMN-EntD and IGMN-IGKE, we also observe that the Causality and Structure types of contradictions take an considerable place. Only with the Entity Distinction type, the total accuracy falls for 4.07%. We must emphasize that the entities in our framework include actions, which are often regarded as relationships in other NLP methods. Thus the Entity Distinction type is relatively comprehensive in our method. Such decline from ablating



...Composite cones, shield volcanoes, cinder cones and supervolcanoes are some of the types of volcanoes formed. Composite cones are steep sided, cone shaped volcanoes that produce explosive eruptions. Shield volcanoes form very large, gently sloped volcanoes with a wide base. Cinder cones are the smallest volcanic landform. They are formed from accumulation of many small fragments of ejected material. A caldera forms when an explosive eruption leaves a large crater when the mountain blows apart...

What do all four types of volcanos have in common?

a. ash b. caldera
c. vent d. magma

GT: **d** BSL: vent
GD: ✗✗✗√ IGMN: d

...The endoplasmic reticulum and Golgi apparatus play a role in producing proteins and moving materials, respectively... The Golgi apparatus is a large organelle that sends proteins and lipids where they need to go. Its like a post office. It receives molecules from the endoplasmic reticulum. It packages and labels the molecules. Then it sends them where they are needed. Some molecules are sent to different parts of the cell. Others are sent to the cell membrane for transport out of the cell. Small bits of membrane pinch off the Golgi apparatus to enclose and transport the proteins and lipids...

What cell structure is surrounded by the endoplasmic reticulum?

a. vacuole b. mitochondria
c. golgi body d. nucleus

GT: **d** BSL: golgi apparatus
GD: ✗✗✗✗ IGMN: d

...The earth is composed of mainly 3 layers: crust, mantle and the core. The crust is the outer layer of the earth... The mantle is the widest section of the earth. The mantle is made up of semi-molten rocks called magma... Core is the innermost layer of the earth. It is further divided into upper core and inner core. Inner core is the center and the hottest part of the earth. It is solid and made up of iron and nickel with temperatures of up to 5500C. It lies between 5150-6370km below earth's crust. Outer core is the layer surrounding the inner core. It is a liquid layer made up of iron and nickel with temperatures similar to inner core. It lies between 2900-5150 km below the earth's crust...

By what letter is the mantle represented in the diagram?

a. A b. W c. H d. P

GT: **c** BSL: W
GD: ✗✗√✗ IGMN: c

...The earth is composed of mainly 3 layers: crust, mantle and the core. The crust is the outer layer of the earth It is a thin layer between 0-33 km thick.... The mantle is the widest section of the earth. The mantle is made up of semi-molten rocks called magma... Core is the innermost layer of the earth. It is further divided into upper core and inner core. Inner core is the center and the hottest part of the earth... Outer core is the layer surrounding the inner core...

Identify V.

a. inner core b. outer core
c. crust d. mantle

GT: **d** BSL: crust
GD: √✗✗✗ IGMN: a

Figure 5: Examples for qualitative analysis, including 2 normal diagram questions and 2 diagram-completion questions. The success and failure cases of the baseline model (BSL), the IGKE module of our method (GD), our whole framework (IGMN) and the ground truth (GT) are compared. The passages beneath the images are selected manually, only for reference. The red color indicates the answers is different from the GT, while the green color means success. BSL (MemN+VQA [16]) generates a shot phrase to find the closest option. GD judges every option and give true ($\surd$) or false ($\times$) suggestions by whether there is a contradiction. IGMN predicts the index of the suggested option.

Causality and Structure types is significant to prove the rationality of the contradiction categories proposed in this paper.

3.5. Qualitative Analysis

Figure 5 shows the qualitative analysis by four examples. In the first example, the baseline fails perhaps because the corresponding essay repeatedly mentions that vents are in many types of volcanoes, but actually vents are not in caldera volcanoes. It observes that our IGKE module provides the correct answer which suggests that the CERGs have the ability to reason throughout the large context. In the second example, the baseline fails possibly because the essay has mentioned that the endoplasmic reticulum works together with the golgi apparatus, but has never mentioned the surrounding relationship. The IGKE module also fails, suggesting that only by CERGs we do not have the ability to recognize the the surrounding relationship from the image, since there is no essential information in the essay. But our final model gives the true answer which demonstrates that the attention and memory networks in our model have the power to correct the mistakes in the Guidance using plenty training data. The comparison between the third and fourth examples shows the limitation of our method. The two questions are the same by human common sense. Our method successes the third example possibly because the blanks are given in order. In the image of the fourth question, the blanks are scrambled so that both of our IGKE and AGJR module fail to provide the correct answer, which shows that it is still difficult for IGMN to comprehend the concepts such as "the innermost" and "outside layer".

The examples further support our belief that it is a promising direction to incorporate the deep neural networks with the semantic analysis to conduct the complex question reasoning, since the graph structure is an effective way to represent the long-term essays while the DNNs are suitable to fuse data of different modalities implicitly.

4. Conclusions

In this paper, we introduce the Instructor Guidance with Memory Networks (IGMN), a novel framework that aims at improving textbook question answering (TQA) task by integrating deep models with semantic analysis. The novelties of our work consist of proposing CERGs and new categories of contradictions to find contradictions by graph matching to generate the Guidance. We further integrate the textual and visual information with the attention-based memory networks, and the corresponding answer can be consequently generated by interpreting the latent reasoning results. Our method has outstanding performance in both the text-only sub-task and the image engaged task of TQA.

Acknowledgements

The work is supported by the National NSF of China (Nos. 61620106010, 61621136008, 61332007, 61571261 and U1611461), Beijing Natural Science Foundation (No. L172037), Tsinghua Tiangong Institute for Intelligent Computing and the NVIDIA NVAIL Program, and partially funded by Microsoft Research Asia and Tsinghua-Intel Joint Research Institute.

References

[1] A. H. Abdulnabi, B. Shuai, S. Winkler, and G. Wang. Episodic CAMN: Contextual attention-based memory networks with iterative feedback for scene labeling. In *Confer-*

ence on Computer Vision and Pattern Recognition (CVPR), pages 5561–5570, 2017.

[2] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra. Vqa: Visual question answering. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015.

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

[4] O. Bakhshandeh and J. Allen. Apples to apples: Learning semantics of common entities through a novel comprehension task. In *ACL*, 2017.

[5] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media Inc.

[6] K. Clark and C. D. Manning. Improving coreference resolution by learning entity-level distributed representations. In *ACL*, 2016.

[7] M.-C. de Marneffe, A. N. Rafferty, and C. D. Manning. Finding contradictions in text. In *ACL*, 2008.

[8] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016.

[9] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *AISTATS*, 2011.

[10] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *CoRR*, abs/1612.00837, 2016.

[11] A. Graves, A. rahman Mohamed, and G. E. Hinton. Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, 2013.

[12] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *CoRR*, abs/1410.5401, 2014.

[13] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1693–1701, 2015.

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–1780, 1997.

[15] A. Kembhavi, M. Salvato, E. Kolve, M. J. Seo, H. Hajishirzi, and A. Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

[16] A. Kembhavi, M. Seo, D. Schwenk, J. Choi, A. Farhadi, and H. Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4999–5007, 2017.

[17] L. Li, B. Qin, and T. Liu. Contradiction detection with contradiction-specific word embedding. *Algorithms*, 10:59, 2017.

[18] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu. Textboxes: A fast text detector with a single deep neural network. In *AAAI*, 2017.

[19] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[20] N. Mrksic, D. Ó. Séaghdha, B. Thomson, M. Gasic, L. M. Rojas-Barahona, P. hao Su, D. Vandyke, T.-H. Wen, and S. J. Young. Counter-fitting word vectors to linguistic constraints. In *HLT-NAACL*, 2016.

[21] B. Pan, H. Li, Z. Zhao, B. Cao, D. Cai, and X. He. Memen: Multi-layer embedding with memory networks for machine comprehension. *CoRR*, abs/1707.09098, 2017.

[22] C. C. Park, B. Kim, and G. Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 895–903, 2017.

[23] M. Ren, R. Kiros, and R. S. Zemel. Exploring models and data for image question answering. In *NIPS*, 2015.

[24] M. J. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. Bidirectional attention flow for machine comprehension. *CoRR*, abs/1611.01603, 2016.

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[26] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng. Parsing with compositional vector grammars. In *ACL*, 2013.

[27] H. Wang, M. Bansal, K. Gimpel, and D. A. McAllester. Machine comprehension with syntax, frames, and semantics. In *ACL*, 2015.

[28] W. Wang, N. Yang, F. Wei, B. Chang, and M. Zhou. Gated self-matching networks for reading comprehension and question answering. In *ACL*, 2017.

[29] J. Weston, S. Chopra, and A. Bordes. Memory networks. *CoRR*, abs/1410.3916, 2014.

[30] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[31] P. Xie and E. P. Xing. A constituent-centric neural architecture for reading comprehension. In *ACL*, 2017.

[32] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *ICML*, 2016.

[33] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, 2016.

[34] Y. Zhu, O. Groth, M. S. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4995–5004, 2016.