Are You Smarter Than A Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension

Aniruddha Kembhavi[†] Minjoon Seo[§*] Dustin Schwenk[†] Jonghyun Choi[†]
Ali Farhadi[†§] Hannaneh Hajishirzi[§]
[†]Allen Institute for Artificial Intelligence, [§]University of Washington
[†]{anik,dustins,jonghyunc,alif}@allenai.org, [§]{minjoon,hannaneh}@washington.edu

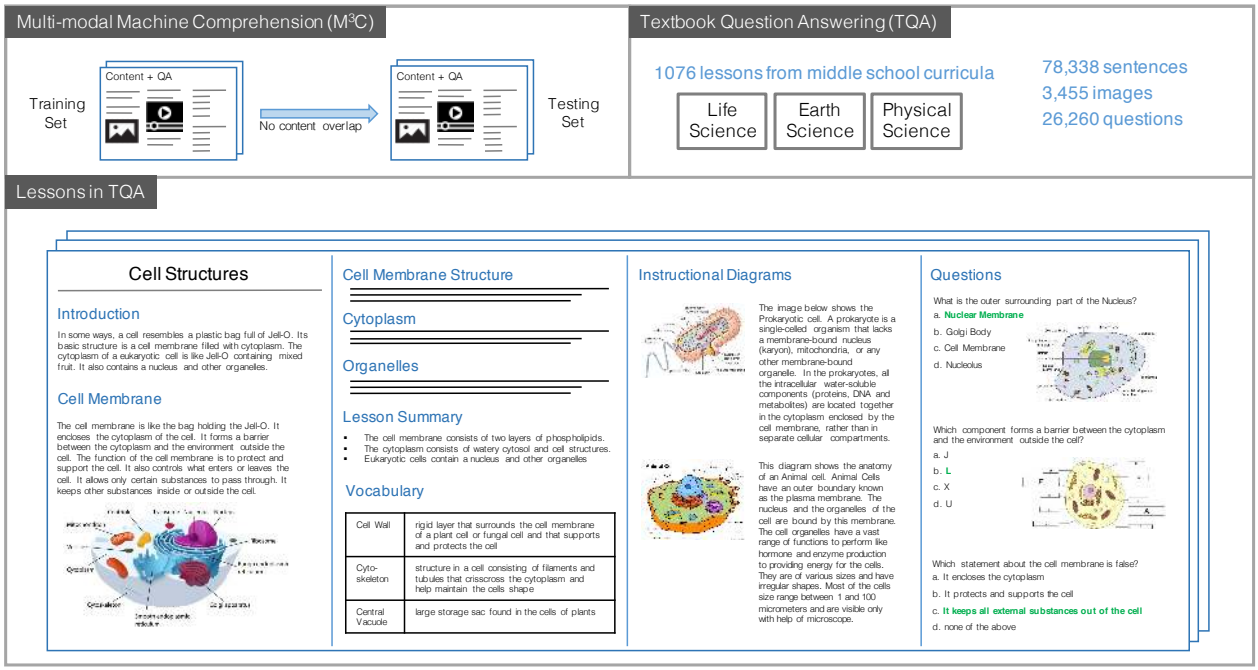


Figure 1. An overview of the Multi-modal Machine Comprehension (M^{3}C) paradigm, statistics of the proposed Textbook Question Answering (TQA) dataset and an illustration of a lesson in it. TQA can be downloaded at http://textbookqa.org .

Abstract

We introduce the task of Multi-Modal Machine Comprehension (M^{3}C), which aims at answering multimodal questions given a context of text, diagrams and images. We present the Textbook Question Answering (TQA) dataset that includes 1,076 lessons and 26,260 multi-modal questions, taken from middle school science curricula. Our analysis shows that a significant portion of questions require complex parsing of the text and the diagrams and reasoning, indicating that our dataset is more complex compared to previous machine comprehension and visual question answering datasets. We extend state-of-the-art methods for textual machine comprehension and visual question answering to the TQA dataset. Our experiments show that these models do not perform well on TQA. The presented dataset opens new challenges for research in question answering and reasoning across multiple modalities.

1. Introduction

Question answering (QA) has been a major research focus of the natural language processing (NLP) community for several years and more recently has also gained significant popularity within the computer vision community.

There have been several QA paradigms in NLP, which can be categorized by the knowledge used to answer questions. This knowledge can range from structured and confined knowledge bases (e.g., Freebase [4, 3]) to unstructured and unbounded natural language form (e.g., documents on the web [24]). A middle ground between these approaches has been the popular paradigm of Machine Comprehension

*The majority of the work was done while the author was interning at the Allen Institute for Artificial Intelligence

(MC) [20, 18], where the knowledge (often referred to as the *context*) is unstructured, and restricted in size to a short set of paragraphs.

Question answering in the vision community, referred to as *Visual Question Answering (VQA)*, has become popular, in part due to the availability of large image-based QA datasets [17, 19, 29, 1, 30, 9]. In a sense, VQA is a machine comprehension task, where the question is in natural language form, and the context is the image.

World knowledge is multi-modal in nature, spread across text documents, images and videos. A system that can answer arbitrary questions about the world must learn to comprehend these multi-modal sources of information. We thus propose the task of Multi-Modal Machine Comprehension (M^3C), an extension of the traditional textual machine comprehension to multi-modal data. In this paradigm, the task is to *read* a multi-modal context along with a multi-modal question and provide an answer, which may also be multi-modal in nature. This is in contrast with the conventional question answering task, in which the context is usually about a single modality (either language or vision).

In contrast to the VQA paradigm, M^3C also has an advantage from a modelling perspective. VQA tasks typically require common sense knowledge to answer many questions, in addition to the image itself. For example, the question *"Does this person have 20/20 vision?"* from the VQA dataset [1] requires the system to detect eye-glasses and then use the common sense that a person with perfect or 20/20 vision would typically not wear eye glasses. This need for common sense makes the QA task more interesting, but also leads to an unbounded knowledge resource. Since automatically acquiring common sense knowledge is a very difficult task (with a large body of ongoing research), it is a common practice to train systems for VQA solely on the training splits of these datasets. The resulting systems can thus only expect to answer questions that require common sense knowledge implicitly contained within the questions in the training splits. The knowledge required for M^3C on the other hand is bounded to the multi-modal context supplied with the question. This makes the knowledge acquisition more manageable and serves as a good test bed for visual and textual reasoning.

Towards this goal, we present the Textbook Question Answering (TQA) dataset drawn from middle school science curricula (Figure 1). The textual and diagrammatic content in middle school science reference fairly complex phenomena that occur in the world [13]. Our analysis in Section 4 shows that parsing this linguistic and visual content is fairly challenging and a significant proportion of questions posed to students at this level require *reasoning*. This makes TQA a good test bed for the M^3C paradigm. TQA consists of 1,076 lessons containing 78,338 sentences and 3,455 images (including diagrams). Each lesson has a set of questions which are answerable using the content taught in the lesson. The TQA dataset has 26,260 questions with 12,567 of them having an accompanying diagram, split into training, validation and test at a lesson level.

We describe the Textbook Question Answering (TQA) dataset in Section 3 and provide an in-depth analysis of the lesson contexts, questions and answer sources in Section 4. We also provide baselines in Section 5 using models that have been proven to work well in other MC and VQA tasks. These models extend attention mechanisms between query and context, where the context (visual and textual) is fit within a memory. Our experiments show that these models do not work very well on TQA. This is presumably due to the following reasons: The length of the context (lessons) is very large and training an attention network (Memory Networks [26]) of this size is non-trivial; there are many different modalities of information that need to be combined into the memory. Most questions cannot be answered by simple lookup, require information from multiple sentences and/or images, and require non-trivial reasoning; Current approaches for multi-hop reasoning work well on synthetic data like bAbI [25], but are hard to train in a general setting such as this dataset. These challenges offered by the TQA dataset make it a valuable resource for the vision and natural language communities, and we encourage other researchers to work on this challenging task. TQA can be downloaded at `http://textbookqa.org`.

2. Background

Visual Question Answering There has been a surge of interest in the field of *language and vision* over the past few years, most notably in the area of visual question answering. This has in part been motivated by the availability of large image and video question answering datasets.

The DAQUAR dataset [16] was one of the earliest question answering datasets in the image domain. Soon after, much larger datasets including COCO-QA [19], FM-IQA [9], Visual Madlibs [29] and VQA [1] were released. Each of these four datasets obtained images from Microsoft COCO dataset [14]. While COCO-QA questions were automatically generated, the remaining datasets used human annotators to write questions. In contrast to our TQA dataset, in all these datasets the question is in a natural language form, and the context is an image. More recently, Zhu *et al*. released the Visual7W dataset [30] which contained multiple choice visual answers in addition to textual answers. While most past works and datasets in the field of question answering in language and vision focused on images, researchers have also made inroads using videos. Tapaswi *et al*. released the Movie-QA dataset [23] which requires the system to analyze clips in the movie to answer questions. They also provide movie-subtitles, plots and scripts as additional information sources.

The presented TQA dataset differs from the above datasets in the following ways. First, the contexts as well as the questions are multi-modal in nature. Second, in contrast to the above VQA paradigm (learn from question-answer pairs and test on question-answer pairs), TQA uses the proposed paradigm of M^3C (read a context and answer questions; learn from context-question-answer tuples and test on context-question-answer tuples). In contrast to the VQA paradigm which often requires unbounded common-sense knowledge to answer many questions, the M^3C paradigm confines the knowledge required to the accompanying context. Another big difference arises from the use of science textbooks and science diagrams in TQA as compared to natural images in past datasets. Science diagrams often represent complex concepts, such as events or systems, that are difficult to portray in a single natural image. Along with the middle school science concepts explained in the lesson text, these images lend themselves more easily to questions that require reasoning. Hence TQA serves as a great QA test bed with confined knowledge acquisition and reasoning.

Early works on visual question answering (VQA) involved encoding the question using a Recurrent Neural Network, encoding the image using a Convolutional Neural Network and combining them to answer the question [1, 17]. Subsequently, attention mechanisms were successfully employed in VQA, whereby either the question in its entirety or the individual words attend to different patches in the image [30, 27, 28]. More recently, [15] employed attention both ways, between the text and the image and showed its benefits. The winner of the recent VQA workshop employed Multimodal Compact Bilinear Pooling [8] at the attention layer instead of the commonly used element wise product/concatenation mechanisms. Our baselines show that networks with standard attention models do not perform very well on the TQA dataset and we discuss the reasons with possible solutions in Section 5.

Machine Comprehension in NLP Akin to the availability of several VQA datasets in computer vision, the NLP community has introduced several machine comprehension (MC) datasets over the past few years. Cloze datasets (where the system is asked to fill in words that have been removed from a passage) including CNN and DailyMail [10] as well as Childrens Book Test [11] are a good proxy to the traditional MC tasks and have the added benefit of being automatically produced. More traditional MC datasets such as MCTest [20] were limited in size, but recently larger ones such as the Stanford Question Answering (SQuAD) dataset have been introduced with 100,000 questions.

Attention mechanisms, largely inspired by Bahdanau *et al*. [2] have become very popular in textual MC systems. There are several variations to using attention including dynamic attention [10, 6] where the attention weights at a time step depend on attention weights at previous time steps. Another popular technique employed is based on Memory Networks [26, 27] with a multi-hop approach, where the attention layer is followed by a query summarization stage and then fed into more rounds of attention on the memory.

The release of the SQuAD dataset has led to a number of new approaches proposed for the task of MC. We extended the approach by Seo *et al*. [21], which currently lies at position 2 on the SQuAD leaderboard, to adapt it to our Multimodal MC task[1]. Our results show that on the text questions, the absolute accuracy is lower than its achieved numbers on the SQuAD dataset. This along with our analysis in Section 4 indicate that the TQA is quite challenging and warrants further research.

3. TQA Dataset

We now describe the Textbook Question Answering dataset and provide an in-depth analysis in Section 4.

3.1. Dataset Structure

The Textbook Question Answering (TQA) dataset is drawn from middle school science curricula. It consists of 1,076 lessons from Life Science, Earth Science and Physical Science textbooks downloaded[2] from `http://www.ck12.org`. This material conforms to national and state curriculum guidelines and is actively being used by teachers and students in the United States and worldwide.

Lessons Figure 1 shows an overview of the dataset. Each lesson consists of **textual content**, in the form of paragraphs of text as well as **visual content**, consisting of diagrams and natural images. Each lesson also comes with a **Vocabulary Section** which provides definitions of scientific concepts introduced in that lesson and a **Lesson Summary** which is typically restricted to five sentences and summarizes the key concepts in that lesson. In total, the 1,076 lessons consist of 78,338 sentences and 3,455 images. In addition, lessons also contain links to online **Instructional Videos** (totalling 2,156 videos across all lessons) which explain concepts with more visual illustrations[3].

Instructional Diagrams We found that textual content in the textbooks was very comprehensive and sufficient to understand the concepts presented in the lesson. However, the textual content and image captions did not comprehensively describe the images presented the lessons. As a result, the lessons were not sufficient to understand the concepts and answer all questions with diagrams. We conjecture that this knowledge gap is filled by teachers in the classrooms, explaining a concept and an accompanying diagram on the

[1] Code available at allenai.github.io/bi-att-flow

[2] All materials from the CK-12 website were downloaded in Aug 2016

[3] Instructional videos are not a part of the TQA dataset. We provide these links as an extension to the dataset to encourage future research in extracting content from instructional videos.

whiteboard. To bridge this gap in the dataset, we added a small set of diagrams (typically between three to five), which we refer to as *Instructional Diagrams*, to lessons in the textbooks that have diagram questions (Section 3.2). We also add rich captions describing the scientific concepts illustrated in the diagram. An example is shown in Figure 1.

Questions Each lesson has a set of multiple choice questions that address concepts taught in that lesson. The number of choices varies from two to seven. TQA has a total of 26,260 questions including 12,567 having an accompanying diagram. We hereby refer to questions with a diagram as *diagram questions*, and ones without as *text questions*.

Dataset Splits TQA is split into a training, validation and test set at lesson level. The training set consists of 666 lessons and 15,154 questions, the validation set consists of 200 lessons and 5,309 questions and the test set consists of 210 lessons and 5,797 questions. On occasions, multiple lessons have an overlap in the concepts they teach. Care has been taken to group these lessons before splitting the data, so as to minimize the concept overlap between data splits.

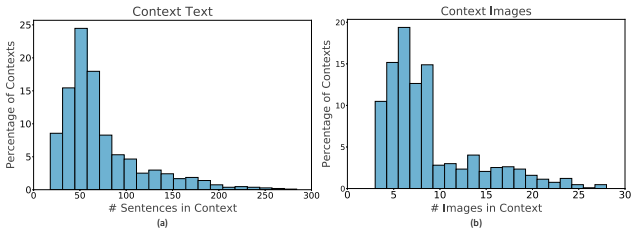


Figure 2. Distribution of textual and visual elements across the lesson contexts in the TQA dataset. (a) Distribution of the number of sentences (b) Distribution of the number of images (including diagrams). Section 4.1 provides a discussion.

3.2. Dataset Curation

The lessons in the TQA dataset are obtained from the Life Science, Earth Science and Physical Science Textbooks and Web Concepts downloaded from the CK-12 website. Lessons contain text, images, links to instructional videos, vocabulary definitions and lesson summaries. Questions are obtained from Workbooks and Quizzes from the website. Additional diagram questions and instructional diagrams are obtained using crowd-sourcing.

Diagram Questions Our initial analysis showed that the number of diagram questions was very small compared to the number of text questions. In part, this is due to the fact that diagram questions are harder to generate. To supplement this set, we obtained a list of scientific topics from each lesson, used these as queries to Google Image Search and downloaded the top results. These were manually filtered down to images that had content similar to the lessons. We thus obtained 2,749 diagrams spread across 85 lessons. Multiple choice questions for these diagrams were then obtained using crowd-sourcing[4]. Each human subject was provided with the full lesson and a diagram and was asked to write down a middle school science question that required the diagram to answer it correctly, and was answerable using the provided lesson.

Instructional Diagrams We obtained a set of instructional diagrams per lesson using the same method as above, de-duplicating diagrams that were already present in the lessons and diagrams that accompanied questions. Rich captions for this set of diagrams were also obtained using crowd-sourcing. Each human subject was provided with examples of rich captions, the lesson and a diagram and was asked to write down rich captions using the vocabulary and scientific concepts explained in the lesson.

4. TQA Analysis

In this section we provide an analysis of the lesson contexts, questions, answers and the information content needed to answer questions in the TQA dataset.

4.1. Lesson Contexts

Figure 2 shows the distribution of the number of sentences and images across the lessons in the dataset. About 50% of lessons have 5-10 images and more than 75% of the lessons have more than 50 sentences. The length of the lessons in TQA is typically higher than past MC datasets such as SQuAD [18], making it difficult to add the entire context into memory and then attending to it. This suggests the need for either an Information Retrieval based preprocessing step or a hierarchical model such as Hierarchical Memory Networks [5]. Furthermore, the multi-modal nature of the contexts in lessons and questions poses new challenges and warrants further research.

4.2. Questions

Text Questions Figure 3(a) shows the distribution of the length of questions in the dataset. This distribution shows that compared to VQA [1], TQA has longer questions (the mode of the distribution here is 8 compared to 5 for VQA). Figure 3(b) shows the distribution of the questions across the *W* categories (what, where, when, who, why, how and which). Interestingly, the *Other* category has a fair number of questions. Further analysis shows that a good fraction of questions written down in standard workbooks are assertive statements as opposed to interrogative statements. This could be another reason why baseline models in Section 5 perform poorly on the dataset.

Diagram Questions The diagrams in the questions of the TQA dataset are similar to the diagrams in the questions of the AI2D dataset presented by Kembhavi *et al*. [13] in

[4]We used MightyAI for all crowd-sourcing needs in this dataset.

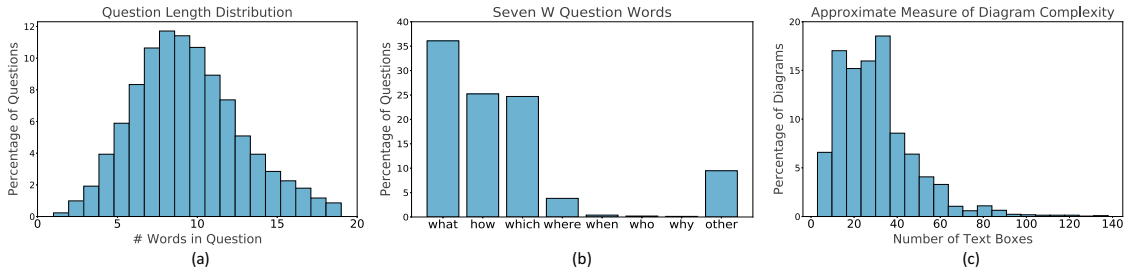


Figure 3. An analysis of questions in the TQA dataset. (a) Distribution of question length (b) Distribution across the *7W* categories (c) Distribution of the number of textboxes for diagrams in questions. Refer to Section 4.2 for further discussion.

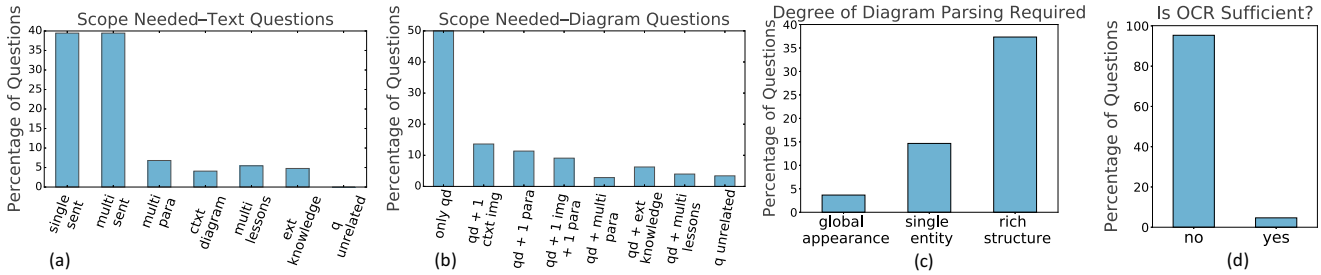


Figure 4. An analysis of the information scope needed to answer questions in the TQA dataset. 'sent', 'para', 'ctxt', 'q', 'qd', 'img' and 'ext' refer to sentence, paragraph, context, question, question in diagram, image and external. (a) Scope needed for text questions. (b) Scope needed for diagram questions. (c) Of the questions that require a diagram, the degree of parsing required. (d) Of the questions that require a diagram, the % of questions that can be answered with the OCR of the diagram alone. Section 4.3 provides more details.

terms of content and complexity. Kembhavi *et al*. propose using diagram parse graphs to represent diagrams and use a hierarchical representation of constituents and relationships. We analysed AI2D and found that there is high correlation between the complexity of a diagram (measured by the number of constituents and relationships in the diagram) and the number of text boxes located in that diagram. Figure 3(c) shows the distribution of the number of text boxes across the diagrams in the questions in the TQA dataset as a proxy to the distribution of diagram complexity. This shows that the diagrams in the questions are quite complex and further analysis below shows that a rich parsing of these diagrams is often needed to answer questions.

4.3. Knowledge Scope to Answer Questions

We also analyze the knowledge scope required to answer questions in the dataset in Figure 4 for each question type. This analysis was performed by human subjects on 250 randomly sampled questions in each type.

Figure 4(a) shows the scope needed for text questions. A significant number of text questions require multiple sentences within a paragraph to be answered correctly, and some questions require information spread across the entire lesson. This is in contrast to past MC datasets like SQuAD [18] where a majority of questions can be answered by 1 sentence. Figure 4(b) shows the scope for diagram questions. Most questions require parsing the question diagram, and of these, a significant number in addition need text and images from the context. Figure 4(c) shows the degree of diagram parsing required to answer questions, given that the diagram is needed. Very few questions can be answered with just a classification of the diagram, and more than 50% need a rich structure to be parsed out of the diagram. Finally, Figure 4(d) shows that fewer than 5% of diagrams can be trivially answered by just the raw OCR text. An example of this case, is where just the correct answer option lies within the text boxes in the image and the wrong options are unrelated to the diagram. This analysis shows that questions in the TQA dataset often require multiple pieces of context information presented in multiple modalities, rendering the dataset challenging.

4.4. Qualitative Examples

True/False Several multiple choice questions in the dataset have just 2 choices: True and False. As one might expect with middle school questions, these are not simple look-up questions but require complex parsing and reasoning. Figure 5 shows 3 examples. The first requires relating *too high* and *below* and also requires parsing multiple sentences. The second requires parsing the flow chart in the diagram and counting the steps. Counting is a notoriously hard task for present day QA systems as has been seen in the VQA dataset [1]. The third requires converting the numerical phrase *2/3* to two thirds as opposed to two and three, and then reasoning that two thirds is more than one-third.

Multiple Choice Figure 6 shows examples of questions

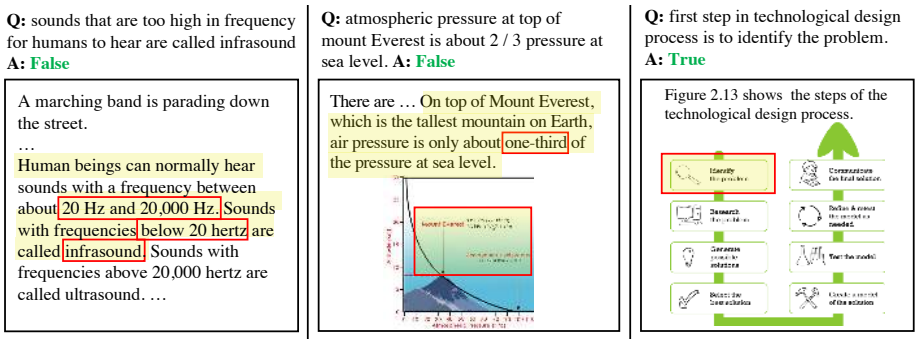


Figure 5. Most True/False questions in TQA require complex parsing and reasoning and are not simple lookup (Section 4.4)

in several interesting categories. (a) requires rich diagram parsing along with a notion of *carries*. (b) multiple sentences and paraphrasing are required. (c) both text and diagram contexts help. (d) multiple sentences are required, and then a notion of order is required. (e) multiple sentences and a notion of *All of the above* are required to pool together results. Interestingly, this is quite a common scenario in the dataset. (f) hypothetical question which are also common to the dataset (g) question requiring analogies. (h) question requiring simple math. It is clear that current state-of-the-art QA models are not designed for such complex tasks, and unsurprisingly, perform very poorly on this dataset.

5. Baselines

We now describe several baseline models and report their performance on Diagram and Text questions in the TQA dataset. These baselines are extensions of the current state-of-the-art models for diagram question answering and textual reading comprehension respectively. We begin by describing the *Text Only Model*. The *Text and Diagram Models* have a very similar architecture and can be considered as extensions to the *Text Only Model*.

5.1. Text Only Model

The *Text Only Model* is an extension to the architecture of Memory Networks [26]. It only considers the textual portions of the questions and lesson contexts. As our analysis in Figure 4 shows, in most cases, this information should be sufficient for answering Text questions, but it is *not* sufficient for answering Diagram questions. The input to the model is a list of paragraphs from the lesson context, the question sentence, and answer choices (2 for True/False questions, 4-7 for Multiple Choice questions). The goal is to output the correct answer among the answer choices.

It is often prohibitive to put all the paragraphs into a GPU's memory. For instance, a single paragraph of 512 words and a batch size of 32 can consume up to 12GB of GPU RAM in a relatively simple architecture. Each lesson often contains more than 1000 words, so a single GPU cannot contain all words (or batch size should decrease, which might degrade performance). A potential solution for handling this issue is using Hierarchical Memory Networks [5]. Here, we choose the most relevant paragraph among the list. We adopt an information retrieval approach: we compute the relevance of each paragraph and the question using tf-idf score of each word, and obtain the paragraph with the highest score of relevance.

Let $\mathbf{M} \in \mathbb{R}^{d \times T}$ represent the embedding of chosen paragraph, where T is the number of words in the paragraph, and d is the size of the embedding for each word. Similarly, let $\mathbf{U} \in \mathbb{R}^{d \times J}$ and $\mathbf{C}_i \in \mathbb{R}^{d \times K}$ represent the embeddings of the question and each choice (i-th choice), respectively. Here, J is the number of the question words, and K is the number of each answer choice sentence. Note that we use padding and masking when necessary to account for different word lengths among the answer choices.

We use Long Short-Term Memory (LSTM) [12] to *embed* each sentence in the paragraph, question, and answer choices. This provides neighboring context to each word. We use ($'$) to indicate that an LSTM has been applied to each modality (e.g. $\mathbf{M}'$ is the LSTM output of $\mathbf{M}$). We then soft-select the word from the paragraph that is most relevant to the question via an attention mechanism. Let $\mathbf{S}_{tj}$ denote the scalar similarity between t-th word of the paragraph and j-th word of the question, computed by

$$\mathbf{S}_{tj} = \mathbf{M'}_{:t}^{\top} \mathbf{U'}_{:j},$$

where $\mathbf{M'}_{:t}$ is the t-th column vector of $\mathbf{M}'$ (corresponding to the LSTM output of t-th word). The attention weight on the paragraph words is obtained by $\mathbf{a} = \mathrm{softmax}(\max_{col} \mathbf{S}) \in \mathbb{R}^T$, where the max is computed over the column of $\mathbf{S}$. Then the attended vector is the weighted sum of the column vectors of $\mathbf{M}'$:

$$\mathbf{m} = \sum_t a_t \mathbf{M'}_{:t} \in \mathbb{R}^d,$$

which can be considered as the predicted answer for the question. We compare the vector with each choice. More concretely, we compute the similarity between the vector and the sum of each $\mathbf{C}'_i$ over the column:

$$\mathbf{r}_i = \mathbf{m}^{\top} \sum_k \mathbf{C'}_{i,:k} \in \mathbb{R}.$$

Then the probability of each choice is the softmax of $\mathbf{r}$, i.e. $\hat{y} = \text{softmax}(\mathbf{r}) \in \mathbb{R}^N$, where N is the number of answer choices. During training, we minimize the negative log probability of the correct answer choice.

5.2. Text + Diagram Models

Text+Diagram Models follow the similar architecture to that of Text Only Model. The only difference is the modality of question and lesson contexts in the memory. We present two diagram baseline models: Text+Image, an extension of state of the art models in the VQA paradigm, and Text+Diagram, an extension to the DSDP-NET model by Kembhavi *et al*. [13] to answer diagram questions.

Text + Image The image is passed through a VGG network [22] (pretrained on Imagenet [7]) and the outputs of the last convolutional layer are added to the memory. The output is a 7-by-7 grid of 512D image patch vectors. As a simple baseline, these 49 vectors can be considered as the context to which the question refers. This is similar to popular models employed in the VQA paradigm (for instance [28]). Our extension involves treating each grid vector in the same way as the LSTM output of the text paragraph in Section 5.1. In order to match the dimension between the LSTM outputs of the paragraph and the grid vectors, we use 2 perceptron layers with `tanh` activation to map each 512D vector to d-dim vector. The transformed vectors are concatenated to the LSTM outputs, so that the question can attend on these image patches in addition to the sentences.

Text + DPG Diagram Parse Graph (DPG) encodes the structured information of the diagram, obtained via parser by [13]. As practiced by the authors, DPG can be translated into factual sentences that describe the graph via several translation rules. For example, if "mouse" object and "cat" object are connected in the DPG, then the translator produces a sentence "mouse is connected to cat.". It is the model's role to associate *connection* to its semantic grounding *eats*. Then these produced sentences can be treated in the same way as the paragraph sentences. The paragraph is initially augmented with these generated sentences; the rest follows the same procedures as in Section 5.1.

5.3. Machine Comprehension Model

We also report the performance of a recently released MC model (BiDAF) [21] on text questions. BiDAF currently ranks second best on the SQuAD leaderboard and has publicly available code. Since BiDAF was originally designed to predict the answer span that lies in the given paragraph (context), we modify its output layer to answer Multiple Choice questions. In particular, the predicted answer span is compared to each answer choice and the one with the highest similarity is chosen as the final answer.

5.4. Baseline Results

Table 1 shows the performance of the four baseline models presented above. Interestingly, both the text models perform very poorly on T/F questions. Most T/F questions in this dataset are not simple lookups but require paraphrasing, multiple sentences, reasoning to be answered correctly (Refer to Figure 5), which standard attention models are not good at. The text models perform better on Multiple Choice questions with roughly 10% improvement over the random baseline. Our analysis in Sec 4.3 and examples in Figure 6 show that many multiple choice questions are complex which explains the poor performance of the baselines.

On Diagram Multiple Choice (MC) questions, we observe that the Text+Image model gives no value beyond the Text only model, but the Text+DPG model performs slightly better than the Text Only Model. This is consistent with the findings in Kembhavi *et al*. in the AI2D dataset [13]. Our analysis in Section 4.3 shows that most diagram questions require a rich diagram parse and often require information across the lesson. Akin to our findings for Text Questions, the standard attention framework in these baselines are unable to handle this level of complexity.

We conjecture that this is mainly due to: (a) contexts in TQA are usually long compared to other MC datasets; (b) fitting multi-modal sources into a single memory introduces new challenges; (c) questions usually require reasoning or show large lexical variations with the context. This introduces new challenges for multi-hop reasoning algorithms beyond synthetic datasets.

Model	Inspired By	Text T/F	Text MC	Diagram MC	All
Random	N/A	50.0	22.7	25.0	28.4
Text Only	MemoryNet [26]	50.2	32.9	29.9	33.8
Text + Image	VQA [1]	N/A	N/A	29.9	33.8
Text + DPG	DSDP-NET [13]	N/A	N/A	31.3	34.6
BiDAF	BiDAF [21]	50.4	32.2	30.1	33.7

Table 1. Baseline Results (% Accuracies). Refer to Section 5.

6. Conclusion

In this paper, we introduce a new task of M^3C as an extension of MC and VQA. We present the TQA dataset as a testbed to evaluate the M^3C task. The TQA dataset consists of 1,076 lessons with 26,260 multi-modal questions. Our experiments show that extensions of the state-of-the-art methods for MC and VQA perform poorly on this dataset, confirming the challenges introduced by this dataset. Future work involves designing systems that can address the M^3C task in the TQA dataset.

Acknowledgements: This work is in part supported by ONR N00014-13-1-0720, NSF IIS-1338054, NSF-1652052, NRI-1637479, NSF IIS-1616112, Allen Distinguished Investigator Award, Google Research Faculty Award, Samsung GRO Award and the Allen Institute for Artificial Intelligence.

(a) Rich Diagram Parsing

Q: This is the long narrow tube that carries food from the pharynx to the stomach.
a. mouth
b. salivary glands
c. liver
d. esophagus

The Components of the Digestive System

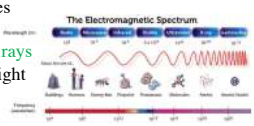
(b) Multiple Sentences

Q: when are most of nadh and fadh2 generated
a) during glycolysis
b) during the krebs cycle
c) during the electron transport chain
d) during cellular respiration

The Krebs Cycle
In the presence of oxygen, under aerobic conditions, pyruvate enters the mitochondria to proceed into the Krebs cycle. The second stage of cellular respiration is the transfer of the energy in pyruvate, which is the energy initially in glucose, into two energy carriers, NADH and FADH2 . A small amount of ATP is also made during this process. This process occurs in a continuous cycle, named after its discover, Hans Krebs. The Krebs cycle uses a 2-carbon molecule (acetyl-CoA) derived from pyruvate and produces carbon dioxide.

(c) Text and Diagram

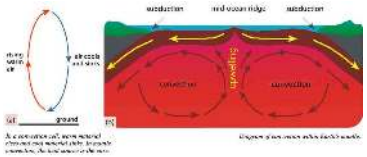
Q: Which of the following choices lists electromagnetic waves from lower to higher frequencies?
a. Radio waves, infrared light, microwaves
b. Ultraviolet light, infrared light, X rays
c. Infrared light, ultraviolet light, gamma rays
d. Visible light, microwaves, ultraviolet light



Light
Radio waves have the longest wavelengths and lowest frequencies of all electromagnetic waves. … On the right side of the diagram are X rays and gamma rays. They have the shortest wavelengths and highest frequencies of all electromagnetic waves.

(d) Order of Events

Q: put in order of how convection currents in the mantle move. i. the material that moved up cools and sinks back down into the mantle. ii. the bottom layer of the mantle material rises and spreads horizontally. iii. the mantle material near the core is heated. iv. the bottom layer of the mantle becomes less dense.
a) iv, iii, ii, i
b) iii, iv, ii, i
c) i, ii, iii, iv
d) iii, i, iv, ii



Heat Flow
Scientists know … 2. Convection: … Convection in the mantle is the same as convection in a pot of water on a stove. …

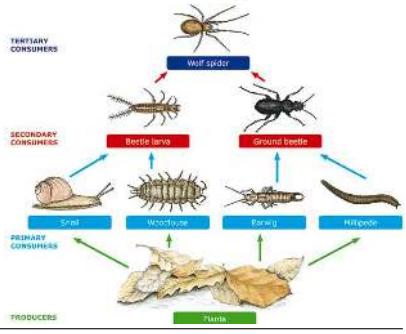
(e) 'N of Above' Answer

Q: What organ(s) do amphibians use to obtain oxygen?
a. gills
b. lungs
c. skin
d. all of the above

Amphibian Skin
… America to poison the tips of their hunting arrows. Amphibian skin contains keratin, a protein that is also found in the outer covering of most other four-legged vertebrates. The keratin in amphibians is not too tough to allow gases and water to pass through their skin. Most amphibians breathe with gills as larvae and with lungs as adults. However, extra oxygen is absorbed through the skin.

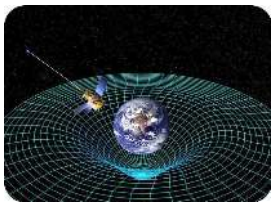
(f) Hypothetical Question

Q: If the population of beetle larva decreases, what happens with the snail population?
a. Decreases
b. Increases
c. Decreases slightly
d. Stays the same



(g) Analogy

Q: Einsteins concept of gravity is similar to what happens when you place a bowling ball on the surface of a trampoline. in this analogy, if the bowling ball represents earth, then the surface of the trampoline represents
a) space-time.
b) earths gravity.
c) earths mass.
d) none of the above



Einstein Explained It All
In the early 1900s, Albert Einstein… showed that gravity is a result of the warping, or curving, of space and time, which made …. relativity.

(h) Simple Math

Q: Assume that a wire has 1.5 ohms of resistance. If the wire is connected to two 1.5-volt batteries, how much current will flow through the wire?
a. 3.0 amps
b. 2.3 amps
c. 2.0 amps
d. 1.0 amps

Ohms Law
Voltage, … Current (amps) = Voltage (volts) Resistance (ohms)
Using Ohms Law to Calculate Current
…If the wire has a resistance of 3 ohms, how much current is flowing through the wire? Current = 12 volts = 4 amps 3 ohms You Try It!

Figure 6. Examples of interesting questions categories in TQA. Green-colored text indicates the correct answer. Red-outlined yellow box illustrates a portion of the lesson textual context useful to answer the question. (Refer to Section 4.4)

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 2, 3, 4, 5, 7

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015. 3

[3] J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, 2013. 1

[4] Q. Cai and A. Yates. Large-scale semantic parsing via schema matching and lexicon extension. In *ACL*, 2013. 1

[5] A. P. S. Chandar, S. Ahn, H. Larochelle, P. Vincent, G. Tesauro, and Y. Bengio. Hierarchical memory networks. *CoRR*, abs/1605.07427, 2016. 4, 6

[6] D. Chen, J. Bolton, and C. D. Manning. A thorough examination of the cnn/daily mail reading comprehension task. In *ACL*, 2016. 3

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 7

[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. 3

[9] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *NIPS*, 2015. 2

[10] K. M. Hermann, T. Kociský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *NIPS*, 2015. 3

[11] F. Hill, A. Bordes, S. Chopra, and J. Weston. The goldilocks principle: Reading children's books with explicit memory representations. In *ICLR*, 2016. 3

[12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997. 6

[13] A. Kembhavi, M. Salvato, E. Kolve, M. J. Seo, H. Hajishirzi, and A. Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 2, 4, 7

[14] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2

[15] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016. 3

[16] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, 2014. 2

[17] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015. 2, 3

[18] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*, 2016. 2, 4, 5

[19] M. Ren, J. R. Kiros, and R. S. Zemel. Exploring models and data for image question answering. In *NIPS*, 2015. 2

[20] M. Richardson, C. J. Burges, and E. Renshaw. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*, 2013. 2, 3

[21] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. Bidirectional attention flow for machine comprehension. In *ICLR*, 2017. 3, 7

[22] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 7

[23] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, 2016. 2

[24] M. Wang, N. A. Smith, and T. Mitamura. What is the jeopardy model? a quasi-synchronous grammar for qa. In *EMNLP-CoNLL*, 2007. 1

[25] J. Weston, A. Bordes, S. Chopra, and T. Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. In *ICLR*, 2016. 2

[26] J. Weston, S. Chopra, and A. Bordes. Memory networks. In *ICLR*, 2015. 2, 3, 6, 7

[27] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *ICML*, 2016. 3

[28] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, 2016. 3, 7

[29] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual madlibs: Fill in the blank description generation and question answering. In *ICCV*, 2015. 2

[30] Y. Zhu, O. Groth, M. S. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, 2016. 2, 3