Technical, Hard and Explainable Question Answering (THE-QA)

Shailaja Sampat
School of Computing, Informatics and Decision Systems Engineering, Arizona State University
ssampa17@asu.edu

Abstract

The ability of an agent to rationally answer questions about a given task is the key measure of its intelligence. While we have obtained phenomenal performance over various language and vision tasks separately, 'Technical, Hard and Explainable Question Answering' (THE-QA) is a new challenging corpus which addresses them jointly. THE-QA is a question answering task involving diagram understanding and reading comprehension. We plan to establish benchmarks over this new corpus using deep learning models guided by knowledge representation methods. The proposed approach will envisage detailed semantic parsing of technical figures and text, which is robust against diverse formats. It will be aided by knowledge acquisition and reasoning module that categorizes different knowledge types, identify sources to acquire that knowledge and perform reasoning to answer the questions correctly. THE-QA data will present a strong challenge to the community for future research and will bridge the gap between state-of-the-art Artificial Intelligence (AI) and 'Human-level' AI.

1 Introduction & Related Work

Data-driven algorithms and neural network architectures have significant contribution in achieving big successes over various language tasks (like question answering, machine translation and text generation) and vision tasks (such as object detection, scene and action recognition). A few tasks have addressed vision and language jointly, such as image captioning, visual question answering and visual relationship extraction, making this an emerging field in ongoing AI research.

However, we are still far from achieving Artificial General Intelligence. The key reason for huge gap between AI and 'general AI' is computers lacking in cognitive ability to acquire new knowledge and do reasoning themselves. This fact motivated the creation of this new challenging corpus, which requires additional level of cognitive ability to reason with respect to the given information, rather than simple lookup from images or text. Through this research, we want to explore the aspect of AI where combined learning and reasoning with traditional data-driven approaches can improve performance.

Visual question answering (VQA) has been explored extensively through VQA, DAQUAR, COCO-VQA datasets, which aims at answering question through image lookup. NLVR, CLEVR, VCR datasets go beyond simple lookups and require some reasoning to answer the questions. Datasets such as Aristo, TQA have been proposed to test the capability of AI models to reason over technical piece of knowledge with provided scientific imagery (if any). Going further, we introduce THE-QA task which puts on additional level of difficulty for reasoning based VQA task- in terms of having a supporting reading unit. Note that, the important trait of this dataset is that an image and corresponding reading passage cannot replace each other (in terms of information they contain) for answering a question. This trait makes THE-QA task little hard as it requires to correctly draw the inferences which is jointly grounded in an image and a reading unit.

2 Research Problem

Understanding complex imagery and answering questions based on it in a given context is a challenge for automation in many real-world tasks. 'Technical, Hard and Explainable Question Answering' (THE-QA) is a new corpus for 'situated question answering' which refers to answering a query by jointly interpreting a question and some environment i.e. an image and a passage in this context. The term *'technical'* corresponds to the use of scientific diagrams in the corpus; the term *'hard'* depicts the possible need of external knowledge and reasoning to answer the questions; and the term *'explainable'* refers to providing explanations for the answers to simulate human like thought process.

2.1 Dataset Creation

Question patterns for THE-QA corpus is inspired by PISA test (Program for International Student Assessment) [OECD, 2016], similar to example provided in Figure 1. This test aims at assessing students on combined understanding of textual and visual inputs as well as general reasoning ability. We plan to compile this dataset consisting of $\sim$5k samples, using template based figure and passage generation by scrapping huge web data and manual formation of QA pairs. Then, we will use several state-of-the art models and mechanisms to establish the baselines and evaluate the performance over this dataset. Also, we propose a modularized 4-step approach of problem solving which is described in the following section;

2.2 Example from THE-QA Dataset

Provided Image

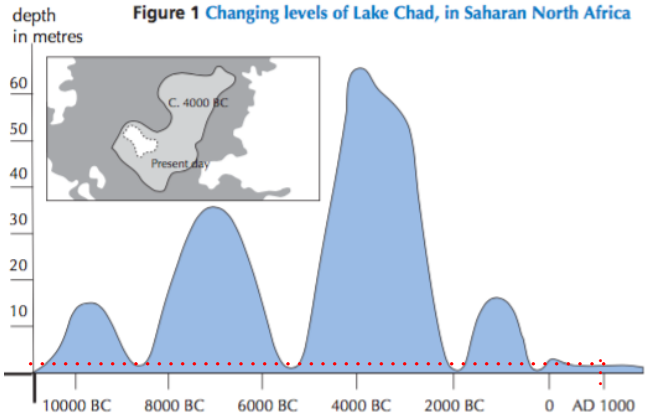

Figure 1: Provided image for a sample question from THE-QA, inspired from PISA test

Provided Reading Passage

Figure 1 shows changing levels of Lake Chad. It disappeared completely in about 20000 BC, during the last ice age. In about 11000 BC, it reappeared. Today, its level is about the same as was in AD 1000.

Question and Answer Choices

What is the depth of Lake Chad today?
A) About 2m. B) About 15m. C) About 50m.
D) It disappeared completely. E) Information not provided.

3 Methodology and Evaluation

Our proposed baseline solver is a modular 4-step method for question answering over THE-QA; First, we perform semantic parsing of imagery and text (passage, questions and choices), then locate the description from passage that is most closely related to the question. Followed by that, we refer to the figure to locate further information as required in question. Finally, we consult back with the question, choices and pick the most likely choice as the final answer.

3.1 Natural Language Understanding

We plan to use state-of the-art machine comprehension tools for the implementation of this task. For explainability of the text understanding modules, we will capture semantics of natural language sentences by converiting them to formal representations using tools like NL2KR (Natural Language to Knowledge Representation) [Baral *et al.*, 2013].

3.2 Image Understanding

Scientific diagram understanding is the most challenging part of this research, especially when figures are unstructured. We will first extract semantically meaningful meaningful constituents for a given question from the image Then we will organize graphical constituents as a parameterized semantic structures and analyze their correlation. We also plan to generate explainable middle level representation of figures using techniques like scene description graphs [Aditya *et al.*, 2015].

3.3 Knowledge Acquisition

We will provide detailed annotations along with the dataset which will incorporate all knowledge required to answer the questions. A starting point in knowledge acquisition can be existing knowledge bases and linguistic ontologies which can be readily used. For the additional requirement of knowledge, we will use advanced Information Extraction tools.

3.4 Reasoning

Addition of a formal reasoning layer to standard statistical machine learning approaches seem to significantly increase the reasoning capability of an agent. With that motivation, we will use Inductive Logic Programming (ILP) [Muggleton, 1991] to automatically learn Answer Set Programs [Gelfond and Lifschitz, 1988] for non-monotonic logical reasoning.

Finally, we will evaluate the proposed 4-step approach through metrics like absolute accuracy, similarity scores, adaptation to noises and model explainability and provide in-depth analysis of relevant ablation studies.

4 Conclusion & Future Work

Deep model-based vision and language understanding algorithms together with a probabilistic knowledge representation and reasoning jointly seems to be a promising solution for answering the hard questions. The findings from this research will advance the development of knowledge-driven question answering, which is essential to overcome the fragility of end-to-end neural models. This work will serve as a foundation for research in AI-based scientific document understanding.

Acknowledgements

This work is being supported by National Science Foundation, Information & Intelligent Systems grant IIS-1816039.

References

[Aditya *et al.*, 2015] Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. From images to sentences through scene description graphs. *arXiv preprint arXiv:1511.03292*, 2015.

[Baral *et al.*, 2013] Chitta Baral, Juraj Dzifcak, Kanchan Kumbhare, and Nguyen H Vo. The nl2kr system. *NLPAR*, 2013, 2013.

[Gelfond and Lifschitz, 1988] Michael Gelfond and Vladimir Lifschitz. The stable model semantics for logic programming. In *Proceedings of JICSLP*, pages 1070–1080. MIT Press, 1988.

[Muggleton, 1991] Stephen Muggleton. Inductive logic programming. *New generation computing*, 8(4):295–318, 1991.

[OECD, 2016] OECD. Pisa: Programme for international student assessment. 2016.