Karlsruhe Institute of Technology

Workshop on Shortcomings in Vision and Language

MoQA – A Multi-Modal Question Answering Architecture

Monica Haurilet, Ziad Al-Halah, Rainer Stiefelhagen

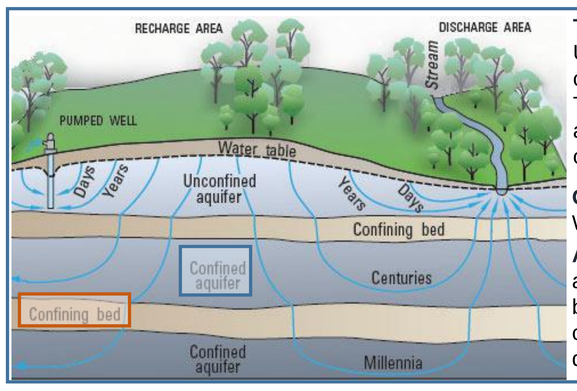
Motivation

Multi-Modal Question Answering

- Task of Answering Questions based on mutliple modalities.
- Modalities: Image, Text, Diagram, etc

Textbook Question Answering (TQA)

- Answering Questions about content in textbooks (Figures and Text)
- Example of Figures: Plots, Tables, Diagrams, ...
- In this work we use diagrams and the surrounding text

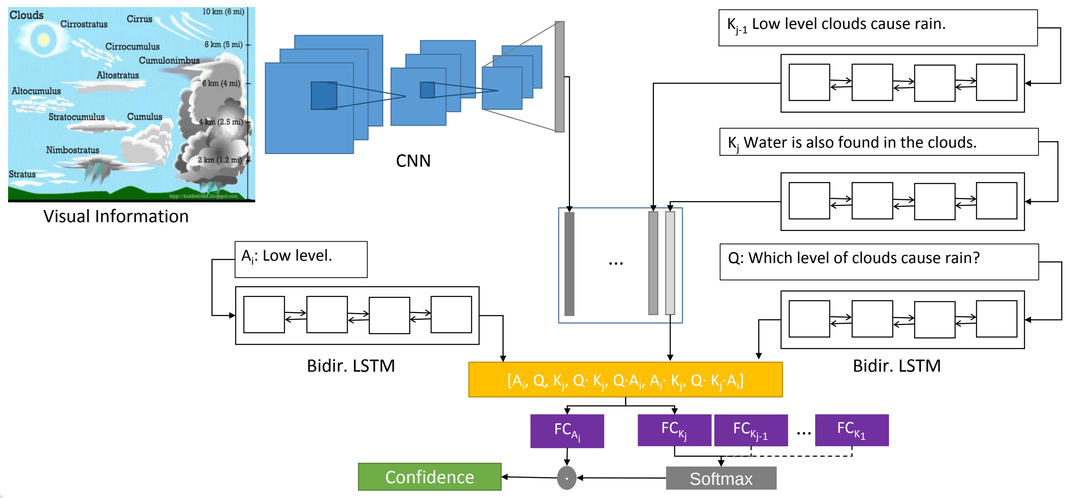


Text
Under the impermeable layer, a thick layer containing water is called confined aquifer. The earth region that supports the confined aquifer is called confining bed. The hole to obtain water in the unconfined aquifer is.…

Question
What layer is underneath the confined aquifer?

Answers
a) Unconfined Aquifer
b) Confined Aquifer
c) Water Table
d) Confining Bed ✓

MoQA



The TQA Dataset

- Questions extracted from **sixth grade** textbooks
- Topics: Life, Earth and Physical Sciences
- 1076 Lessons
- 78K sentences
- 13K **Text Questions** splitted into true false and multiple choice
- 13K **Diagram Questions** with multiple choice questions about diagram and text

Evaluation

Validation Results

- Improvement in T/F Quest.
- Infersent trained for infering relations betw. sentences

Text-only Model won the TQA Challenge 2017:

- **1st place** in the text track
- **2nd place** in the diagram track

	T/F	MC	Text	Diag.
Random	50.1	22.9	33.6	25.0
MemN + VQA	50.5	31.1	38.7	31.8
MemN + VQA + HT	50.3	28.1	36.9	29.8
MemN + DPG	50.5	30.1	38.7	32.8
BiDAF + DPG	50.4	30.5	38.7	32.7
Challenge	–	–	45.6	35.9
IGMN	**57.4**	**40.0**	**46.9**	**36.4**
Ours [InferSent]	61.9	36.2	46.4	33.4
Ours [SkipThought]	60.2	36.4	45.6	34.0

Accuracy on the validation set

Approach

- Per lesson there are over 70 sentences in text-form for each question
- Only a small subset of the text relevant to answer the question
- We select K sentences and feed them to our neural network
- The neural network picks the relevant sentence using an attention module
- Image representation is included in the same embedding space as sentences
- Question, answers and knowledge representation fused using concatenation

Representation Types

Graph-only

- Similar to the text-only model
- Instead of sentences we use edges
- Represted by text+location of node pair

Graph Baseline

- Select most similar node to question
- Choose most similar answer to nearest node

Image-only

- Only global image representation
- No attention is used

Modality	#S	#N	Diag.
Image-only	–	–	33.2
Text-only	3	–	33.8
	4	–	33.9
	5	–	33.4
Graph-baseline	–	1	29.2
	–	2	28.3
	–	4	25.8
Graph-only	–	4	33.3
Text+Image	4	–	34.0

Supporting Sentences

Assumption

∃a set of sentences K_j that are able to verify the correctness of (Q, A_i)

Method

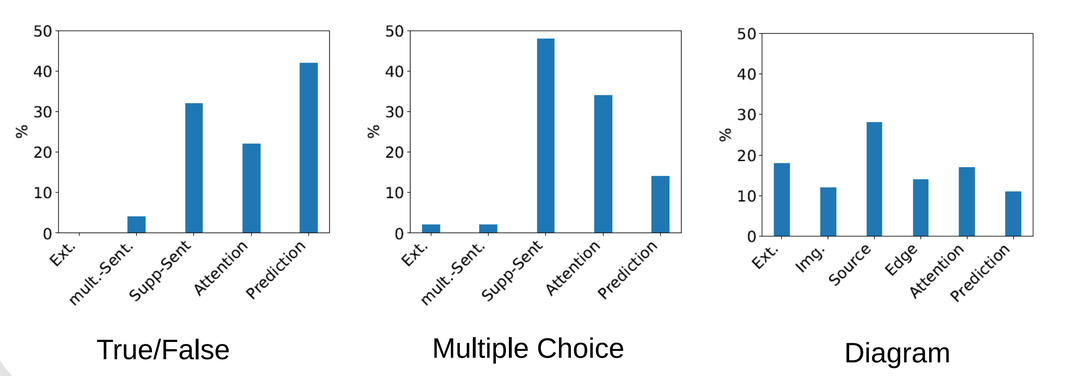
Select the top-k most similar sentences to the question
Use these sentences to verify (Q, A_i)

Question
Human actions that increase the risk of soil loss include

Answers
a) Logging ✓
b) Terracing
c) Tree planting
d) No tree planting

Supporting Sentences
1) Human actions that can increase soil erosion are described below
2) Other human actions that put soil at risk include logging, mining, and construction ✓
3) Show how farming practices can increase soil erosion
4) There are several other ways to help prevent soil loss

In-depth Analysis



Contact

Email haurilet@kit.edu
Website https://cvhci.anthropomatik.kit.edu/~mhaurile