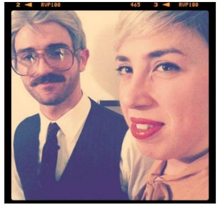
Visual question answering

State of the art

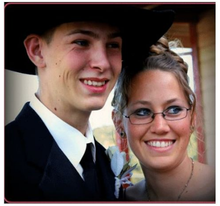
What is Visual Question Answering (VQA)?

Who is wearing glasses?

man

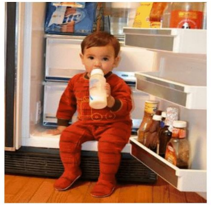


woman

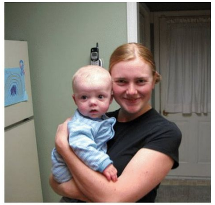


Where is the child sitting?

fridge

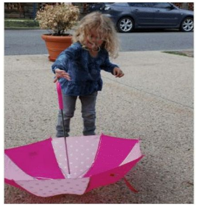


arms

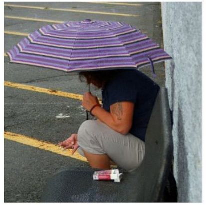


Is the umbrella upside down?

yes

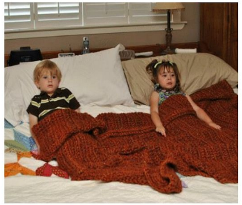


no

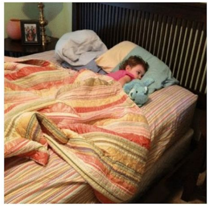


How many children are in the bed?

2



1



Related topics

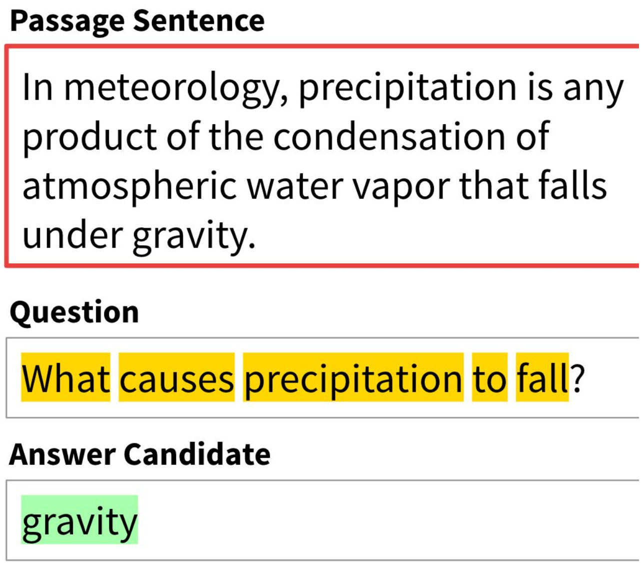
Image captioning

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

Textual question answering



Passage Sentence

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

Question

What causes precipitation to fall?

Answer Candidate

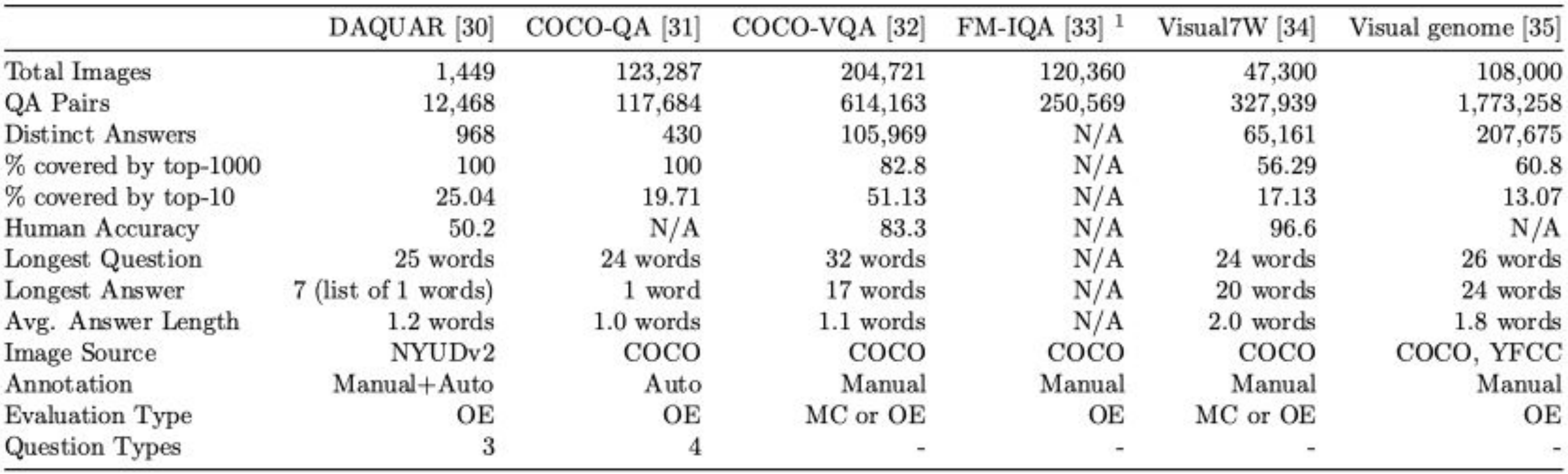
gravity

Overview

- stretches the **natural language processing** and **computer vision** barriers
- can be considered for **Visual Turing Test** for image understanding
- implies deep understanding of scene of images and the relation between objects

1. Object recognition - What is in the image?
2. Object detection - Are there any cats in the image?
3. Attribute classification - What color is the cat?
4. Scene classification - Is it sunny?
5. Counting - How many cats are in the image?
6. Relation - What is between the cat and the sofa?

Datasets

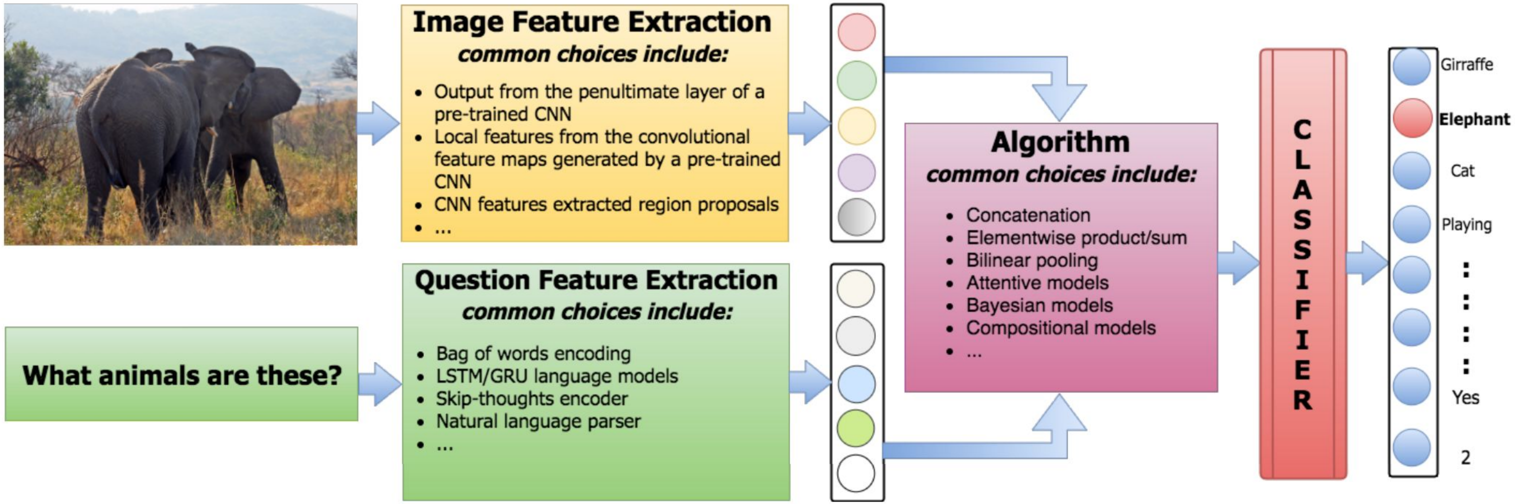


	DAQUAR [30]	COCO-QA [31]	COCO-VQA [32]	FM-IQA [33] [1]	Visual7W [34]	Visual genome [35]
Total Images	1,449	123,287	204,721	120,360	47,300	108,000
QA Pairs	12,468	117,684	614,163	250,569	327,939	1,773,258
Distinct Answers	968	430	105,969	N/A	65,161	207,675
% covered by top-1000	100	100	82.8	N/A	56.29	60.8
% covered by top-10	25.04	19.71	51.13	N/A	17.13	13.07
Human Accuracy	50.2	N/A	83.3	N/A	96.6	N/A
Longest Question	25 words	24 words	32 words	N/A	24 words	26 words
Longest Answer	7 (list of 1 words)	1 word	17 words	N/A	20 words	24 words
Avg. Answer Length	1.2 words	1.0 words	1.1 words	N/A	2.0 words	1.8 words
Image Source	NYUDv2	COCO	COCO	COCO	COCO	COCO, YFCC
Annotation	Manual+Auto	Auto	Manual	Manual	Manual	Manual
Evaluation Type	OE	OE	MC or OE	OE	MC or OE	OE
Question Types	3	4	-	-	-	-

Evaluation metrics

1. Single word answers (yes/no, multiple choice)
    a. accuracy or precision/recall
2. Open ended answers
    a. Wu-Palmer Similarity: semantic meaning
    b. Some datasets have multiple answers to the same question (again accuracy + some fuzziness)
    c. # of manual labelers that had the same answer (usually 3)
    d. manual evaluation

Methods



What animals are these?

Image Feature Extraction

common choices include:

- Output from the penultimate layer of a pre-trained CNN
- Local features from the convolutional feature maps generated by a pre-trained CNN
- CNN features extracted region proposals
- ...

Question Feature Extraction

common choices include:

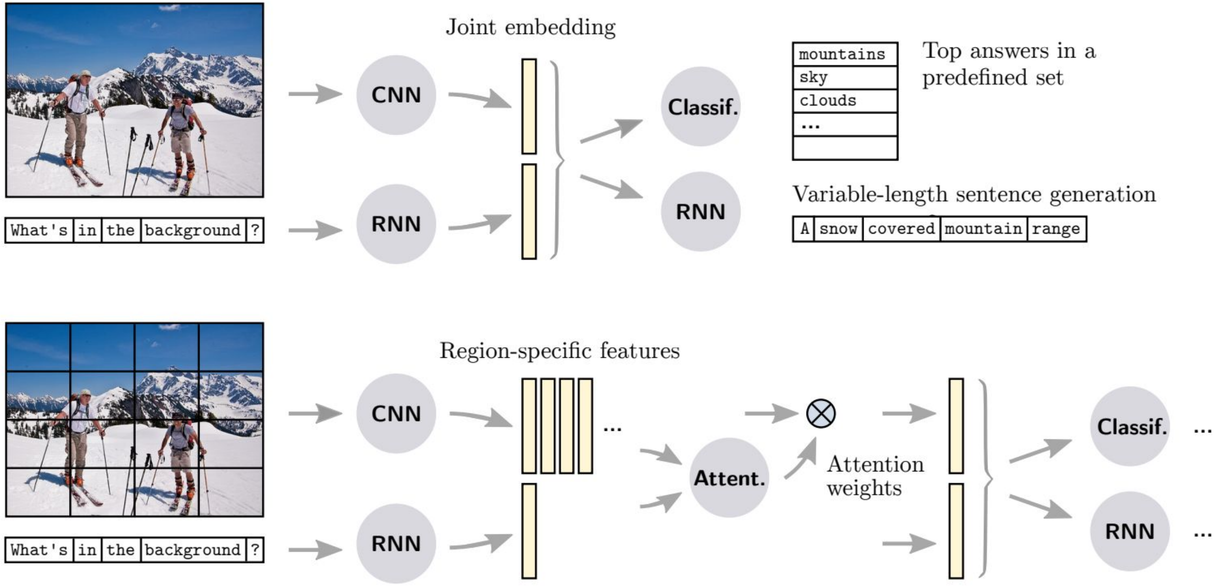
- Bag of words encoding
- LSTM/GRU language models
- Skip-thoughts encoder
- Natural language parser
- ...

Algorithm

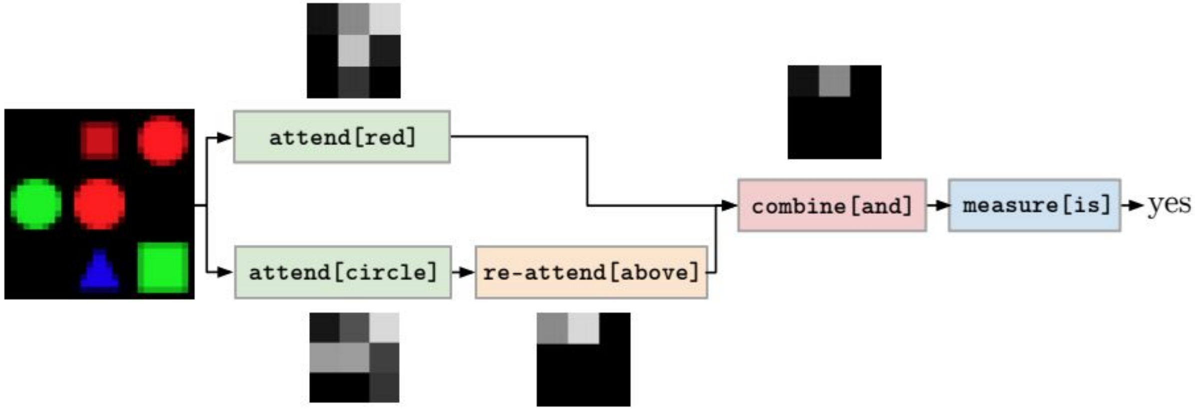
common choices include:

- Concatenation
- Elementwise product/sum
- Bilinear pooling
- Attentive models
- Bayesian models
- Compositional models
- ...

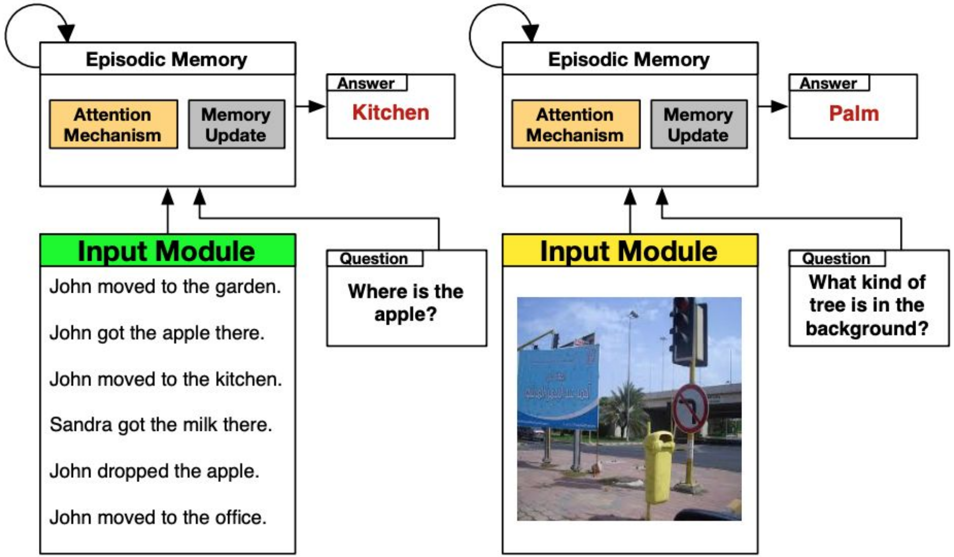
Joint Embeddings + Attention based


Joint embedding
CNN
RNN
Classif.
RNN
What's in the background ?
mountains
sky
clouds
...
Top answers in a predefined set
Variable-length sentence generation
A snow covered mountain range
Region-specific features
CNN
RNN
...
Attent.
Attention weights
Classif.
RNN
...
...
What's in the background ?

Neural Module Networks

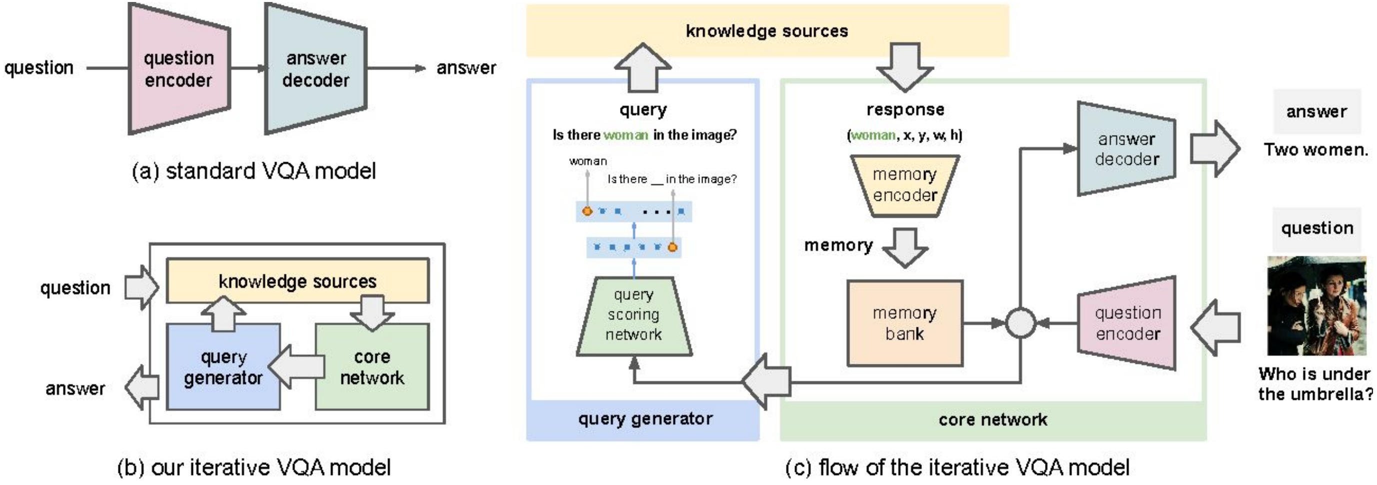


Dynamic memory networks



(a) Text Question-Answering (b) Visual Question-Answering

Knowledge base augmentation



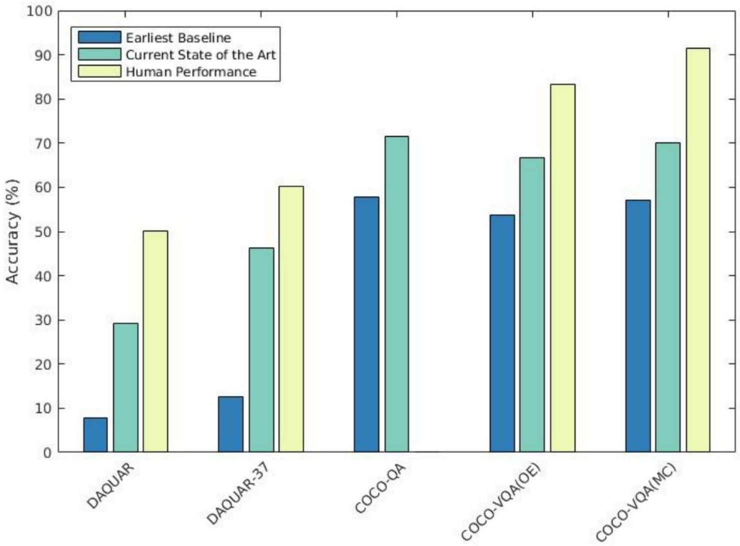
(a) standard VQA model

(b) our iterative VQA model

(c) flow of the iterative VQA model

Challenges

- still the human results are far better than machine results
- bias in question - rephrasing questions provide very different results
- hard to define evaluation metrics
- common sense is hard to extract and to be fed to neural networks
- extracting spatial relationships between entities is a very complex task


Earliest Baseline
Current State of the Art
Human Performance
Accuracy (%)
DAQUAR
DAQUAR-37
COCO-QA
COCO-VQA(OE)
COCO-VQA(MC)