Are You Talking to a Machine?

Dataset and Methods for Multilingual Image Question Answering

Haoyuan Gao; Junhua Mao; Jie Zhou; Zhiheng Huang; Lei Wang; Wei Xu

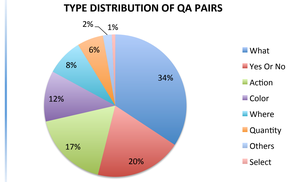
Abstract

Children learn to use language by associating what they hear with what they see. It is more nature for human to get the information about what they see by asking a question. We present the *mQA* model, which is able to answer questions about the content of an image. The answer can be a sentence, a phrase or a single word. We construct a Freestyle Multilingual Image Question Answering (FM-IQA) dataset to train. We construct a Freestyle Multilingual Image Question Answering (FM-IQA) dataset to train and evaluate our *mQA* model. We propose strategies to monitor the quality of this evaluation process. The experiments show that in 64.7% of cases, the human judges cannot distinguish our model from humans.
The project page: http://idl.baidu.com/FM-IQA.html, or scan the QR code on the right

Data Set

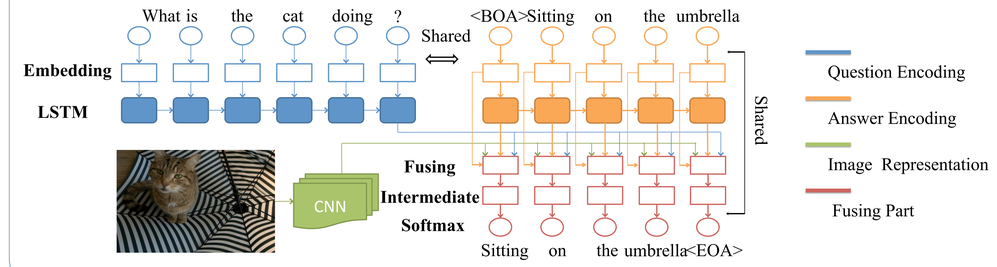
Images : More than 150,000 images From MS-COCO Dataset;
QA-Pairs : More than 300,000 QA pairs in total (Chinese);
FM-IQA : Release parts of QA types;
Types of Questions：

- “What”: questions about the attributes and features of the object.
- “Yes Or No”: questions that you can answer with Yes or No.
- “Action”: questions about the action and behavior of the subject.
- “Color”: questions about the color of the object.
- “Quantity” : questions about the quantity and number of the object.
- “Where”: questions about the location of the object.
- “Select”: Selective questions.
- Other intriguing question.



Samples of Data Set：

The Multimodal QA (mQA) Model



Experiment

The Visual Turing Test:

- 3000 QA-pairs for images
- 12 People for Test

The Score of the Generated Answer:

- “0” : Totally wrong
- “1” : Partially correct
- “2” : Perfectly correct

	Visual Turing Test			Human Rated Scores			
	Pass	Fail	Pass Rate(%)	2	1	0	Average Score
Human	948	52	94.8	927	64	9	1.918
blind-QA	340	660	34	-	-	-	-
mQA	647	353	64.7	628	198	174	1.454

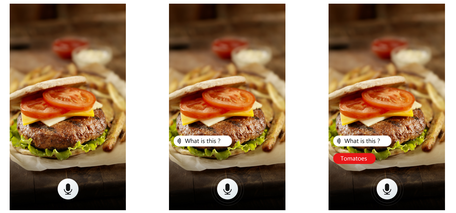
Sample Answers Randomly :

Failed examples :

Interesting Work

Generate Question and Answer:

New Interaction Method:



Discussion

Conclusion:

- We present the mQA model, which is able to give a sentence or a phrase as the answer to a freestyle question for an image.
- We construct a Freestyle Multilingual Image Question Answering (FM-IQA) dataset containing over 310,000 question-answer pairs.
- We evaluate our method through a real Turing Test.

future work:

- we will try to address these issues by incorporating more visual and linguistic information (e.g. using object detection or using attention models).