

Are You Talking to a Machine?

Dataset and Methods for Multilingual Image Question Answering

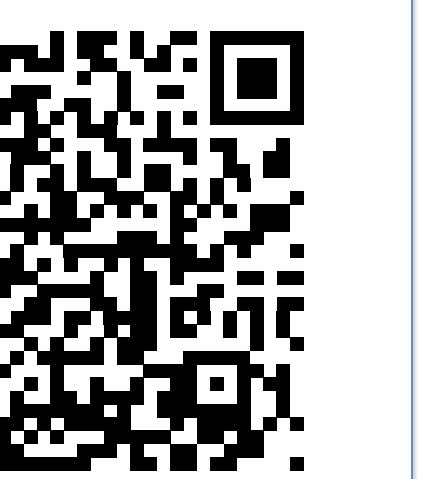
Haoyuan Gao; Junhua Mao; Jie Zhou; Zhiheng Huang; Lei Wang; Wei Xu



Abstract

Children learn to use language by associating what they hear with what they see. It is more natural for humans 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;

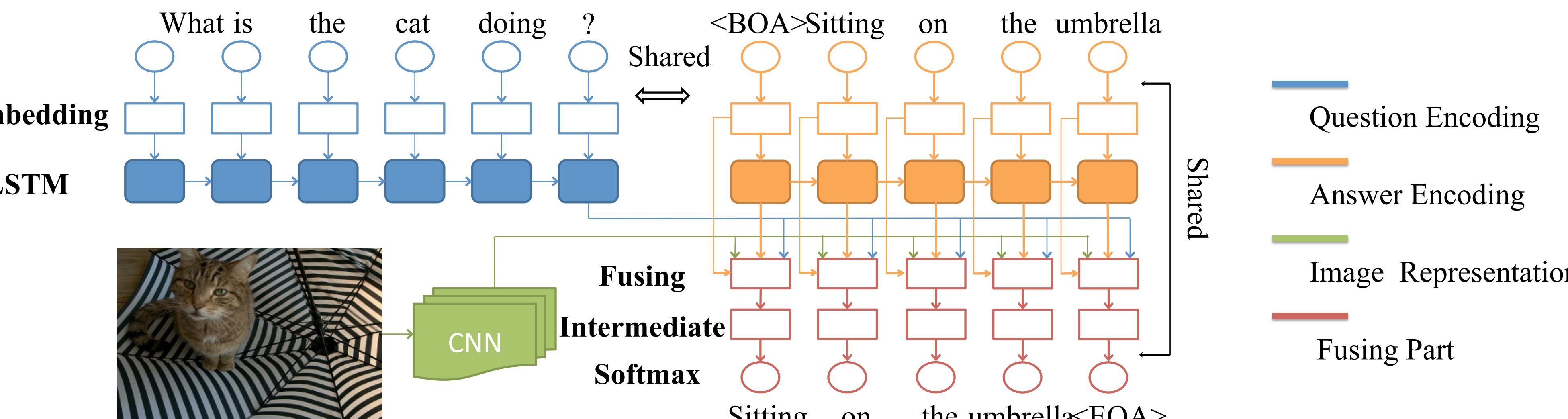
Types of Questions:

- "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:

	Image	GT Question	GT Answer
Image		戴帽子的男孩在干什么? What is the boy in green cap doing?	他在玩滑板。 He is playing skateboard.
GT Question		图片中有人吗? Is there any person in the image?	有。 Yes.
GT Answer		电脑在老人的左面还是右面? Is the computer on the right hand or left hand side of the gentleman?	右面。 Right side.
Image		飞盘是什么颜色? What is the color of the frisbee?	黄色。 Yellow.
GT Question		公交车停在那干吗? Why does the bus park there?	准备维修。 Preparing for repair.
GT Answer		这些人在做什么? What are the people doing?	打雨伞步行。 Walking with umbrellas.
Image		房间里的沙发是什么质地的? What is the texture of the sofa in the room?	布艺。 Cloth.
GT Question		这个蛋糕是几层的? How many layers are there for the cake?	六层。 Six.
GT Answer		手机，鼠标，电脑混放表示什么? What does it indicate when the phone, mouse and laptop are placed together?	主人困了，睡着了。 Their owner is tired and sleeping.

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

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

Sample Answers Randomly :

Image	Question	Answer
	公共汽车是什么颜色的? What is the color of the bus?	公共汽车是红色的。 The bus is red.
	黄色的是什么? What is yellow?	香蕉。 Bananas.
	草地上除了人以外还有什么动物? What is there on the grass, except the person?	羊。 Sheep.
	观察一下说出食物里任意一种蔬菜的名字? Please look carefully and tell me what is the name of the vegetables in the plate?	西兰花。 Broccoli.

Failed examples :

	帅哥在干什么? What is the handsome boy doing?		盘子里有什么水果? Which fruit is there in the plate?		公交车停在那干吗? Why does the bus park there?
	在抓飞盘。 Trying to catch the frisbee.		苹果和橙子。 Apples and oranges.		准备维修。 Preparing for repair.

Interesting Work

Generate Question and Answer:

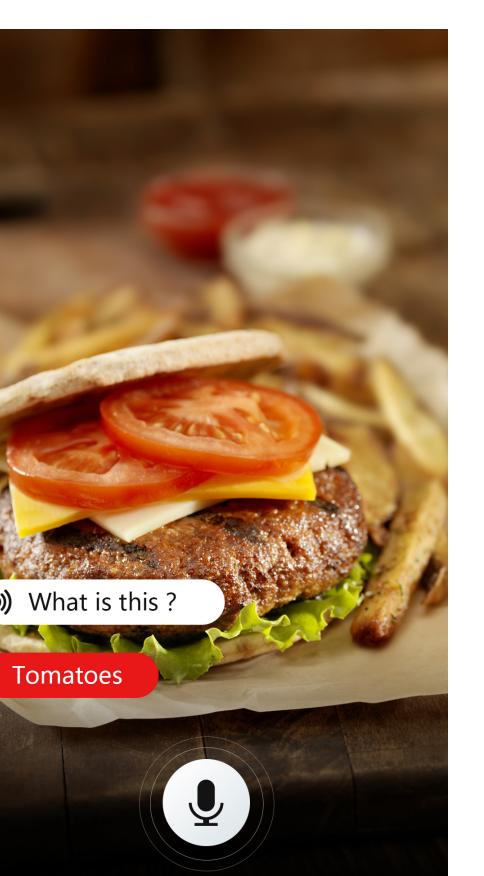


电脑在哪里?
Where is the computer?
在桌子上。
On the desk.



这个人打网球么?
Is this guy playing tennis?
是的。
Yes.

New Interaction Method:



Capture
Ask by voice
Generate Answer
Speak by TTS

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).