Data Science and Intelligent Systems

Visual Question Answering

* **What is VQA**

For a very brief introduction, VQA works as: given an image and a question in natural language, the task is to provide an accurate natural language answer.

It means the whole process contain three parts:

1. Give a image.
2. Give a question.
3. Return the answer.

图形用户界面, 应用程序

描述已自动生成

Figure 1. How VQA works

For most of the VQA problem I’ve seen, the computer receive an image and a textual description about the question. Then the computer will give the correct answer, usually a few words or few choices. A related task is to fill in the blank about the description about a image.

The question and answers very. But usually in the following catalog:

文本

描述已自动生成

Figure 2. Catalog of tasks and questions

* **What Can VAQ do?**

During the very long research, I find that VAQ can’t be a very useful tool. The application mainly in the following fields:

1. **Help blind individuals:**

To help blind users to communicate through pictures. For an example, if one want to know what is on an art work or what is on signs the program can help to read it. But I think VQA can only be used in some supportive tasks. If put into tasks like navigation, it can be dangerous.

1. **Online business**

Its most likely for online shopping. It attracts customers of online shopping sites by giving satisfying results for their search queries. Like now what Amazon’s AI customer service do. What makes it different is VQA can give answers about a item depends on the picture.

1. **Help to learn**

To allow learners engaged in educational services to interact with images. It will mainly for young kids or language learner. It can be interactive by VQA.

1. **Supervisory using**

To help the analysts in surveillance data. It can be applied in security use. Like if you install a surveillance before your door, you can ask the program when and where if someone pass the door or some suspects passed by.

* **How does it works?**

I’m not familiar with deep learning and machine learning. But for most of the case I find, such following steps will the task use:

1. To process images for the neutral network. Here we use CNN.
2. To convert natural language questions to be processable. LSTM is the most used method. (For here I don’t really understand)
3. Joint feature representation, which means to process both features from images and text at the same time.
4. Once the joint feature is determined, then the network can run and answer can be generated.

* **Why I choose this topic?**

I haven’t learnt python or deep learning before. So I think it is a good opportunity to try it. Besides, Although the VQA is not a very exciting topic, the algorithm used in the method seems interesting. For most of the cases, joint feature representation is used. I think it is a very interesting method and I want to figure it out. So I can also learn about deep learning.

* **references:**

[1] *Antol\_VQA\_Visual\_Question\_ICCV\_2015\_paper*

<https://arxiv.org/abs/1512.02167>

[2] *Visual Question Answering: a Survey*

<https://blog.paperspace.com/introduction-to-visual-question-answering/>

[3] *Where to Look: Focus Regions for Visual Question Answering*

https://openaccess.thecvf.com/content\_cvpr\_2016/html/Shih\_Where\_to\_Look\_CVPR\_2016\_paper.html

[4] *Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge* https://openaccess.thecvf.com/content\_cvpr\_2018/html/Teney\_Tips\_and\_Tricks\_CVPR\_2018\_paper.html