Abstract

VQA refers to a technology that searches for elements such as people or things in pictures and generates related questions and answers. It requires the machine to analyze the picture by itself, then extract and classify the elements in the picture, and finally match the keywords in the question. The difficulty of this technique is to analyze the picture and how to ensure the accuracy of the answer to the question.

Introduction

With the popularity of AI equipment and the progress of deep learning in recent years, the correlation between computer vision and natural language processing has become a major research direction. Developers let the machine recognize the surrounding environment and make autonomous judgments and inferences. This technology can provide convenience to the visually impaired because the machine can feed back the recognition result to the user through voice or other methods.

A VQA system takes as input an image and a free-form, open-ended, naturallanguage question about the image and produces a naturallanguage answer as the output.

Datasets

1. COCO-QA

In COCO-QA [1], QA pairs are created for images using a Natural Language Processing

(NLP) algorithm that derives them from the COCO image captions. COCO-QA contains 78,736 training and

38,948 testing QA pairs. Most questions ask about the object in the image (69.84%), with

the other questions being about color (16.59%), counting (7.47%) and location (6.10%).

2. DAQUAR

The first significant VQA dataset was the DAtaset for QUestion Answering on Real-world images (DAQUAR) [2]. It contains 6794 training and 5674 test question-answer pairs, based on images from the NYU-Depth V2 Dataset. That means about 9 pairs per image on average.

Although it is a great initiative, the NYU dataset contains only indoor scenes with, sometimes, lightning conditions that make it difficult to answer the questions. In fact, evaluation on humans shows an accuracy of 50.2%.

1. The VQA Dataset

Compared to other datasets, the VQA dataset [3] is relatively larger. In addition to 204,721 images from the COCO dataset, it includes 50,000 abstract cartoon images. There are three questions per image and ten answers per question, that is over 760K questions with around 10M answers. To achieve this, a team of Amazon Mechanical Turk workers generated the questions and another team wrote the answers.

For the multiple-choice mode, they create 18 candidate answers (correct and incorrect) per question:

The Correct Answer: The most common answer given by the ten annotators.

Plausible Answers: Three answers collected from annotators without looking at the image.

Popular Answers: The top ten most popular answers in the dataset (e.g. “yes”, “no”, “2”, “1”, “white”, “3”, “red”, “blue”, “4”, “green” for real images)

Random Answers: Randomly selected correct answers for other questions.

Methods

Baselines

Baseline [4] methods help determine the difficulty of a dataset, and establish the minimal level

of performance that a more sophisticated algorithms should exceed.

Approaches based on attention

The goal of the attention-based approaches [5] is to set the focus of the algorithm on the most relevant parts of the input.

Bayesian approaches

The idea behind bayesian approaches [6] is to model co-occurrence statistics of both the question and the image features, as a way of inferring relationships between questions and images.

Conclusion

The future of VQA will be to create larger, more diverse data sets that will provide more accurate answers to questions.This allows the system to be applied to real life situations, such as autonomous driving and travel for the visually impaired.These all require a VQA system that is accurate and intelligent enough.

VQA is an important foundation for computer vision and natural language processing, but currently it can only recognize objects within a database.

We believe that VQA should be an essential part of any visual image system, so a VQA system that can answer any question will be the future direction of artificial intelligence.

References

[1]. M. Ren, R. Kiros, and R. Zemel, “Exploring models and data for image question

answering,” in Advances in Neural Information Processing Systems (NIPS), 2015.

[2]. M. Malinowski and M. Fritz, “A multi-world approach to question answering about realworld scenes based on uncertain input,” in Advances in Neural Information Processing

Systems (NIPS), 2014.

[3]. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “VQA:

Visual question answering,” in The IEEE International Conference on Computer Vision

(ICCV), 2015.

[4][5][6]. Kushal Kafle and Christopher Kanan, “Visual Question Answering:

Datasets, Algorithms, and Future Challenges”, 2017