Project 1

Jiachen Xu email:ryanxu@bu.edu

VQA is a multi-disciplinary AI task which combines advanced techniques of computer vision and natural language processing to build a system that can answer a query about an image.

I think VQA can be used in the following scenes.

1. Image retrieval.
   When we need to find a particular image among a great number of images, it can be time-consuming for us to look at the pictures one by one. So, in that case, we can just describe the feature of the image that we want to get. Then we will get it easily. For example, if we want to find all the images of cat in a dataset, we can simply ask 'Is it a cat?' to all images in the dataset.
2. Aided-navigation for blind individuals
   For blind people, they cannot see their surroundings and their life can be very inconvenient when there is no one to guide them. But by using VQA technology, the blind people can ask for what obstacles are in their way, and a captioning system can describe the scene in front of the blind people. Besides they could use VQA to query the image to get more insight about the scene.
3. Automatic querying of surveillance video
   When a police officer wants to find a criminal, sometimes it's hard to search for a guy from maybe a ten hour-long surveillance video. However, by using VQA technology, he can describe the clothes color and the skin color or any other details of the criminal and can precisely find the location of the criminal.

1.VQA Dataset

There are many different datasets that have been proposed for VQA.

(1) DAQUAR(Dataset for Question Answering on Real-world images): This is the earliest VQA data set but also the smallest. This dataset is of poor quality. Images are disorganized and the resolution is low. Besides, there are clear grammatical errors in the questions and answers.

(2) COCO-QA: This dataset is made by COCO dataset, then using NLP algorithm to generate questions and answers. The dataset contains 78,736 pieces of training data and 38,948 pieces of test data. As for the questions raised, 69.84% were about the target, 16.59% were about the color, 7.74% were about the count and 6.10% were about the position. All of the answers were single words, and there were only 435 unique answers. The biggest problem of COCO-QA dataset is that all QA (question-answer) is obtained by NLP algorithm, but the problem is that it can't deal with complex sentences, which leads to some grammatical errors in the question. Another problem is that the question is only designed for the above four aspects.

(3) The VQA Dataset: This dataset is widely used in model evaluation, but the problem with this data set is that many questions are highly similar and unified due to language bias. In addition, since many questions are subjective, there is a certain directivity in the questioning process.
(4) FM-IQA(The Freestyle Multilingual Image Question Answering): The questions and answers of this data set are also done manually. The questions and answers are in Chinese and then translated into English. The difference with the previous dataset is that the answer of this dataset can be a sentence.
(5) Visual Genome: This data set is larger than the previous VQA dataset. The questions of this dataset is 6W which means What, Where, How, When, Who, and Why. The diversity of answers in this dataset is significantly better than that in previous datasets. In addition, there is no question in the form of is or isn't.
(6) Visual7W: This dataset is an extension of the last dataset. 7W refers to What, Where, How, When, Who, Why, and Which.
(7) Shapes: The dataset includes questions about attributes, relationships, and the location of shapes. And all the questions are answered with only "yes or no". Images of this dataset is composed of polygons with different shapes and colors.

2. Evaluation Metrics

There are four different evaluation metrics proposed for VQA

(1) Simple Accuracy
Positives: Very simple to evaluate and interpret. Works well for small number of unique answers
Shortcomings: Both minor and major errors are penalized equally. Can lead to explosion in number of unique answers, especially with presence of phrasal or sentence answers.

(2) Modified WUPS
Positives: More forgiving to simple variations and errors. Does not require exact match. Easy to evaluate with simple script.
Shortcomings: Generates high scores for answers that are lexically related but have diametrically opposite meaning. Cannot be used for phrasal or sentence answers.

(3) Consensus Metric
Positives: Common variances of same answer could be captured. Easy to evaluate after collecting consensus data.
Shortcomings: Can allow for some questions having two correct answers. Expensive to collect ground truth. Difficulty due to lack of consensus

(4) Manual Evaluation
Positives: Variances to same answer is easily captured. Can work equally well for single word as well as phrase or sentence answers.
Shortcomings: Can introduce subjective opinion of individual annotators. Very expensive to setup and slow to evaluate, especially for larger datasets.

3.VQA Models

There are also many different VQA algorithms. But all of the existing methods consist of 1) extracting image features (image featurization), 2) extracting question features (question featurization), and 3) an algorithm that combines these features to produce an answer.

There are different ways to combine image features and question features. Thus, many different models have been proposed:

(1) Attention Based Models: The positives of this kind of model is that the attention mechanism allows us to find the area we want to focus on more quickly.
(2) Bilinear Pooling Methods: Combining the image and question features using bilinear pooling or related schemes in a neural network framework.
(3) Baseline model: Combining the image and question features using simple mechanisms, e.g., concatenation, elementwise multiplication, or elementwise addition, and then giving them to a linear classifier or a neural network.
(4) Bayesian and Question-Aware Model: Using Bayesian models that exploit the underlying relationships between question-image-answer feature distributions.
(5) Compositional VQA Models: These models break the VQA task into a series of sub-tasks, and using different sub-networks to deal with each sub-tasks.

I actually find a javascript demo of VQA model trained on easy-VQA dataset. This dataset is similar with Shapes dataset. They are both consist of polygons with different shapes and colors, and the answers to the questions include the color of polygons besides “yes” or “no”.
https://easy-vqa-demo.victorzhou.com/

A **yellow**, **circle** shape.

Want a different image?

Random Image

What is the color of the shape?

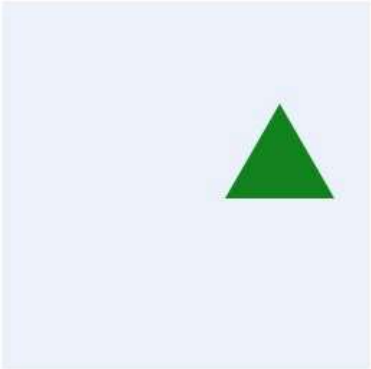
Want a different question?

Random Question

Predict

Prediction: **yellow**

The Image



A **green**, **triangle** shape.

Want a different image?

Random Image

The Question

Is a rectangle present?

Want a different question?

Random Question

Predict

Prediction: **no**

To improve the VQA technology, I think we should build a larger and less biased dataset. And future datasets need more nuanced analysis for benchmarking. The current datasets I found dealing with questions mainly on the targets' color, number and position. Maybe it can be more comprehensive.

I also find an VQA application called CognitiveCam: CognitiveCam is a Visual Question Answering application designed for helping the visually impaired. A user can ask any question pointing towards the object, the app will then attempt to provide the relevant answer. The application is successful enough to identify everyday objects, read text and identify age and gender of a human.

It's definitely a wonderful application, but this cam can only sense the surroundings in front of the user, what if change it to a wide-angle camera. The system can get the whole environment around the user and will tell the more information to the user.

I also think the application can directly speak to users if they find anything unfit to the environment. For example, if cogitivecam find some obstacles like stones in a blind people's way, it would be better if cognitivecam just speak out "There are some stones in front of you, be careful!" Rather than waiting for users to ask questions. Which might cause some danger.

Besides, I think this application can attach some hardware to it, for example, when the cognitivecam find some obstacles in front of the blind people and told them there some obstacles, the blind people can make a command like "Clean it up." Then the mechanical arm will clean up the obstacles so that the blind people can walk through it safely. And when blind people want to get something like a glass of water, he asks where it the water, and the system told them. But it's hard to get that by themselves just according to the description of the system. In such condition, they can also ask the system to bring the water for them, then the mechanical arm will deliver the water to users' hands which can be far more convenient.

Reference

[1] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," arXiv preprint arXiv:1512.02167, 2015.
[2] K. Kafle and C. Kanan, "Answer-type prediction for visual question answering," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[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] 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.
[5] K. Kafle and C. Kanan, \Answer-type prediction for visual question answering," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[6] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[7] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in Conference on Empirical Methods on Natural Language Processing (EMNLP), 2016.
[8] J.-H. Kim, K.-W. On, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," arXiv preprint arXiv:1610.04325, 2016.
[9] K. Kafle, C. Kanan. "Visual Question Answering: Datasets, Algorithms, and Future Challenges." Computer Vision and Image Understanding, 2017