VQA: Visual Question Answering

1 Citation

Antol, Stanislaw, et al. ”Vqa: Visual question answering.” Proceedings of the IEEE international conference on computer vision. 2015.

https://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Antol_VQA_Visual_Question_ICCV_2015_paper.pdf

2 Abstract

In Visual Question Answering, a computer is presented with an image and natural language question about the image. The computer is expected to produce a natural language answer in response. We provide a dataset with 250K images, 760K questions, and 10M answers. We've also evaluated some baseline models.

3 Introduction

A good AI-complete task forces a computer to demonstrate multi-modal knowledge and produces answers that can be easily evaluated quantitatively.

We think Visual Question Answering (VQA) is a good AI-complete task. For example, showing a picture of a boy in glasses and asking ”does this person have 20/20 vision?” requires natural language processing, computer vision, knowledge representation, and commonsense reasoning skills. The performance metric is simply the fraction of questions that the AI answers correctly.

We make a VQA dataset where we take 200K images from the MS COCO dataset and create 50K ”abstract” images (i.e. images of scenes composed from clipart). Each image has three questions, each of which was answered by 10 subjects.

We've provided this dataset publicly, provided some baseline models, and we are going to set up a yearly competition and conference around this problem.

4 Related Work

Other VQA datasets are small or have a limited set of answers or objects in images. Text-based QA struggles to find a frame of reference to ask a question about. In VQA, this frame is obviously the image. There are models that caption images, but the quality of a caption is hard to quantify.

5 VQA Dataset Collection

We take 200K images from MS COCO because they have good diversity. Some researchers may not interested in solving the vision piece and may want to focus on the NLP or commonsense reasoning piece, so we also provide 50K abstract scene dataset which are images where we've put together a cartoon scene using some a large set of clipart.

MS COCO has 5 sentences to go with each image, so we collect these sentences for the abstract scenes as well using crowdsourcing.

To crowdsource questions about the images, we asked subjects to give us a question about the image that would stump a smart robot. We told subjects that it should not be possible to answer the question without seeing the image. For each question, we collected answers from 10 subjects.

At test-time, the model can either provide an open answer or select an answer from a list of possibilities. The accuracy metric is summing the following over all questions posed to the computer:

$$\min\left(\frac{\text{num humans who gave same answer}}{3}, 1\right)$$

To create the multiple choice list, we combine the correct answer with plausible, popular, and random answers. To get plausible answers, we crowdsourced answers to questions without showing the image (or use a nearest neighbor in a bag-of-words).

6 VQA Dataset Analysis

We get a lot of "What is...", "Is the...", and "How many..." questions, but there is a good diversity. Most questions have between 4-10 words.

Most answers are a single word. About 40% of questions are yes/no questions. Of these yes/no questions, yes is the correct answer about 58% of the time. For "how many" questions, 26% of the time, the answer is 2.

When we posed questions to subjects, we asked how confident they were in their response. Subjects are usually confident. Subjects tend to be in agreement with each other more often on yes/no questions, probably because we don't correct for synonyms in other types of questions.

We asked subjects how old a human would have to be to answer the question. Most questions are at the level of toddlers, young kids, and older kids.

We also tried providing an image caption (no image) and question to subjects. They do a little better than if we only gave them the question, but they do a LOT better when we give them the image and the question.

7 VQA Baseline and Methods

If you pick randomly, accuracy is 0.12%. If you always answer "yes", you get 30%. If you pick the most popular answer based on type of question, you get 36%. Nearest neighbor gets 40%.

We modeled the problem as 1000-way classification (over 1000 possible answers) and tried solving with two models: a multi-layer perceptron and an LSTM. We looked at the question-only, question and caption, and question and image input cases. We use bag-of-words of the 1000 most popular caption words to represent the text in the first two cases (we one-hot-encode for the LSTM) and we use VGG's last hidden layer as the image representation. Our best model is the LSTM that processes both the image and question. It gets 54% accuracy and performs at the level of a 4.5 year old.

8 Conclusion and Discussion

We provide a dataset, give some baseline models, and set up a conference/ competition/evaluation-server for Visual Question Answering. Our questions are open-ended, so this is a good AI-complete problem.