Visual Question Answering
Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, Lawrence Zitnick, Devi Parikh
(ICCV 2015)

Yanan Liu
04/12/2018
Outline

• Problem Statement
• Motivations
• The VQA Dataset
• Discussion of the Dataset
• Baselines
• Conclusion
What is VQA

Given an image and a natural language question about the image, the task is to provide a natural language answer.
what are in the picture

Predicted top-5 answers with confidence:
elephants 99.133%
elephant 0.707%
people 0.017%
trees 0.010%
zebras 0.008%

what is in the picture

Predicted top-5 answers with confidence:
elephant 89.543%
elephants 9.952%
horse 0.126%
man 0.053%
people 0.027%
what are the things in the picture doing

Predicted top-5 answers with confidence:
walking  61.979%
standing  26.749%
eating    4.166%
playing   3.278%
drinking  0.572%

is it rainy

Predicted top-5 answers with confidence:
no        82.649%
yes       17.351%
day       0.000%
cloudy    0.000%
0          0.000%
Motivations

• Applications: helping the visually impaired

• “AI completeness” (image captioning can be too simple)

• A task has to:
  • Be open-ended
  • Require multi-modal knowledge (CV, NLP, KR)
  • Suitable for quantitative evaluation
Multi-modal knowledge

**Question:** What can the red object on the ground be used for?

**Answer:** Firefighting

**Support Fact:** Fire hydrant can be used for fighting fires.
Quantitative Evaluation Metric

• Open Ended Answering V.S. Multiple Choice

• Suitable for Self-Evaluation:
  • Multiple choice
  • Simple answer
    • Yes or no
    • Number

How many cats are there?

(a) One
(b) Two
(c) Three
(d) Four
VQA Dataset

• 250k Images (200k MS COCO + 50k abstract images)
• Each image with 3 questions (750k questions)
• Each question with ~10 answers (10M answers)
• Train/Val/Test split: 2:1:2
Distribution of Questions

what is (13.84)
what color (08.98)
what kind (02.49)
what are (02.32)
what type (01.78)
is the (10.16)
is this (08.26)
how many (10.28)
are (07.57)
does (02.75)
where (02.90)
is there (03.60)
why (01.20)
which (01.21)
do (01.15)
what does (01.12)
what time (00.67)
who (00.77)
what sport (00.81)
what animal (00.53)
Distribution of Answers

• Most answers are short:
  • ~90% one word, ~6% two words, ~2.5% three words

• Brevity of the answers doesn’t mean the problem is easy:
  • 23,234 unique one-word answers

• ~40% are “yes”/”no” answers, among which ~56% are “yes”
• ~13% percent are numbers, among which “2” is the most popular
Distribution of Answers
More Discussion

- Subject has high confidence about their answers
- Inter-human agreement on the answers is high
- Common sense knowledge is needed
- Image is needed

![Bar charts for Real Images and Abstract Scenes showing the percentage of questions with different levels of agreement based on average confidence.](chart.png)
Baselines

• Randomly choose an answer from top 1k answers: 0.12%
• Always select the most popular answer (“yes”): 29.72%
• Pick the most popular answer per question type: 36.18%

• Classification
  • Output: 1 of the top 1k answers (cover 82.67% answers in the train+val set)
  • Model: MLP, LSTM
  • Input: Question(bag of words), Caption(bag of words), Image(last hidden layer of VGGNet), Q+I, Q+C, Q+I+C
## Baselines

\[
\text{min}\left( \frac{\text{# humans that provided that answer}}{3}, 1 \right)
\]

<table>
<thead>
<tr>
<th></th>
<th>Open-Answer</th>
<th></th>
<th>Multiple-Choice</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Yes/No</td>
<td>Number</td>
<td>Other</td>
</tr>
<tr>
<td>Question</td>
<td>48.09</td>
<td>75.66</td>
<td>36.70</td>
<td>27.14</td>
</tr>
<tr>
<td>Image</td>
<td>28.13</td>
<td>64.01</td>
<td>00.42</td>
<td>03.77</td>
</tr>
<tr>
<td>Q+I</td>
<td>52.64</td>
<td>75.55</td>
<td>33.67</td>
<td>37.37</td>
</tr>
<tr>
<td>LSTM Q</td>
<td>48.76</td>
<td>78.20</td>
<td>35.68</td>
<td>26.59</td>
</tr>
<tr>
<td>LSTM Q+I</td>
<td>53.74</td>
<td>78.94</td>
<td>35.24</td>
<td>36.42</td>
</tr>
<tr>
<td>Caption</td>
<td>26.70</td>
<td>65.50</td>
<td>02.03</td>
<td>03.86</td>
</tr>
<tr>
<td>Q+C</td>
<td>54.70</td>
<td>75.82</td>
<td>40.12</td>
<td>42.56</td>
</tr>
</tbody>
</table>
Conclusion

• The task of VQA is introduced.
  • Open-ended
  • Multi-modal knowledge
  • Suitable for self-evaluation

• A VQA dataset is provided.

• Annual challenge, workshop, etc.
Further Reading