Agenda

- Problem Description
- Related Works
- Dataset Overview
- Dataset Analysis
- Model Description
- Results
- Discussion & Questions
Purpose:

• Given an:
  • Image
  • Question (Free-Form and Open Ended)
Purpose:

What is the mustache made of?
Consist of...

Graph Credit: Zitnick
Consist of...

- Object Detection & Recognition
Consist of...

- Activity Recognition
Consist of...

- Knowledge-Based Reasoning
- Commonsense Reasoning
Related Works

• Those which works on synthetic small datasets
• Those which generates template answers to template questions
  • Fixed vocabulary of objects, attributes and relationship between objects
  • Few type of questions
• Text Based Question Answering
• Image&Video Tagging, Image&Video Captioning
• Those which are not related at all
Dataset (Overview)

- 204,721 images from MS COCO Dataset
  - Diverse and Complex Scenes (hence diverse questions)
- 50,000 images from Abstract Scenes Dataset
  - Abstract HandMade Scenes
  - Realistic
  - No need to parse images, concern only on high level reasoning
- 3 Questions per image (~760K Questions)
- 10 Answers per question with confidence of answerer (~10M Answers)
- 5 captions per image
- 4.7 Human Years!
Dataset (Images)

- 204,721 images from MS COCO Dataset
  - 123,287 training/validation
  - 81,434 test

- Questions are also product of images
  - Diverse, comprehensive and interesting questions.
Dataset(Images)

- 50,000 images from Abstract Scenes Dataset
  - No Noise, recognizer
  - Genders, Ages, Races, Expressions, 100 objects, 31 animals at various poses
  - Imititates the real images than previous works

20K/10K/20K => train/validation/test
Dataset(Questions)

- require low-level vision:
  - What color is the cat?
  - How many chairs are there?
- require commonsense knowledge:
  - What sound the animal in picture makes?
- require image!
  - What is the mustache made of?

Image Credit: Antol
Dataset (Questions)

Stump a smart robot! Ask a question about this image that a human can answer, but a smart robot probably can’t!

Updated instructions: Please read carefully

We have built a smart robot. It understands a lot about images. It can recognize and name all the objects, it knows where the objects are, it can recognize the scene (e.g., kitchen, beach), people’s expressions and poses, and properties of objects (e.g., the color of objects, their texture, etc). Your task is to stump this smart robot!

Ask a question about this image that this SMART robot probably can not answer, but any human can easily answer while looking at the image.

IMPORTANT: The question should be about this image. That is, the human should need the image to be able to answer the question -- the human should not be able to answer the question without looking at the image.

Your work will get rejected if you do not follow the instructions below:

- Do not repeat questions. Do not ask the same questions or the same questions with minor variations over and over again across images. Think of a new question each time specific to each image.

- Each question should be a single question. Do not ask questions that have multiple parts or multiple sub-questions in them.

- Do not ask generic questions that can be asked of many other images. Ask questions specific to each image.

Please ask a question about this image that a human can answer *if* looking at the image (and not otherwise), but would stump this smart robot:

Q1: Write your question here to stump this smart robot.
Do These Questions Need Commonsense to Answer?

We will present you with a series of questions about images. For each question, please indicate whether or not you think the question requires commonsense in order to answer. A question requires commonsense to answer if answering the question requires some knowledge beyond what is directly shown in the image. Some examples are provided below.

Q: How many calories are in this pizza?

To answer this question, is commonsense required?
1. yes
2. no
Spectrum

3-4 (15.3%)
Is that a bird in the sky?
What color is the shoe?
How many zebras are there?
Is there food on the table?
Is this man wearing shoes?

5-8 (39.7%)
How many pizzas are shown?
What are the sheep eating?
What color is his hair?
What sport is being played?
Name one ingredient in the skillet.

9-12 (28.4%)
Where was this picture taken?
What ceremony does the cake commemorate?
Are these boats too tall to fit under the bridge?
What is the name of the white shape under the batter?
Is this at the stadium?

13-17 (11.2%)
Is he likely to get mugged if he walked down a dark alleyway like this?
Is this a vegetarian meal?
What type of beverage is in the glass?
Can you name the performer in the purple costume?
Besides these humans, what other animals eat here?

18+ (5.5%)
What type of architecture is this?
Is this a Flemish bricklaying pattern?
How many calories are in this pizza?
What government document is needed to partake in this activity?
What is the make and model of this vehicle?

Slide Credits: Agrawal
VQA Age

- Average “age of questions” = 8.98 years.
- Our model =* 4.74 years old!

* age as estimated by untrained crowd-sourced workers
Dataset(Questions)

Human subjects were told:

“We have built a smart robot. It understands a lot about images. It can recognize and name all the objects, it knows where the objects are, it can recognize the scene (e.g., kitchen, beach), people’s expressions and poses, and properties of objects (e.g., color of objects, their texture). Your task is to stump this smart robot!”

Please ask questions that require images to answer.
Dataset (Questions)

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>what is</td>
<td>13.84</td>
</tr>
<tr>
<td>what color</td>
<td>08.98</td>
</tr>
<tr>
<td>what kind</td>
<td>02.49</td>
</tr>
<tr>
<td>what are</td>
<td>02.32</td>
</tr>
<tr>
<td>what type</td>
<td>01.78</td>
</tr>
<tr>
<td>is this</td>
<td>10.16</td>
</tr>
<tr>
<td>is this</td>
<td>08.26</td>
</tr>
<tr>
<td>how many</td>
<td>10.28</td>
</tr>
<tr>
<td>are</td>
<td>07.57</td>
</tr>
<tr>
<td>does</td>
<td>02.75</td>
</tr>
<tr>
<td>where</td>
<td>02.90</td>
</tr>
<tr>
<td>is there</td>
<td>03.60</td>
</tr>
<tr>
<td>why</td>
<td>01.20</td>
</tr>
<tr>
<td>which</td>
<td>01.21</td>
</tr>
<tr>
<td>do</td>
<td>01.15</td>
</tr>
<tr>
<td>what does</td>
<td>01.12</td>
</tr>
<tr>
<td>what time</td>
<td>00.67</td>
</tr>
<tr>
<td>who</td>
<td>00.77</td>
</tr>
<tr>
<td>what sport</td>
<td>00.81</td>
</tr>
<tr>
<td>what animal</td>
<td>00.53</td>
</tr>
<tr>
<td>what brand</td>
<td>00.36</td>
</tr>
</tbody>
</table>
Dataset(Answers)

• Many questions can be answered «yes» or «no»
  • Is this task that easy?

• Other questions can be answered by short phrases

• Subjects who hadn’t asked the questions were told:
  • Don’t use conversational language
  • Use Brief phrases
  • Provide your confidence on answering question

• To avoid conflicts on correct answer, 10 subjects answered
Dataset(Answers)

• For open-ended questions
  • Accuracy = \( \min \left( \frac{\text{#humans that provided that answer}}{3}, 3 \right) \)

• Word2Vec does not work here:
  • as white-red, left-right or young-old are similar in terms of their vector representations
Dataset (Answers)

• **Multiple Choice Answers** :
  - **Corrects**: most common out of 10 answers for image
  - **Plausible**: answered without seeing image -> to take image into account
  - **Popular**: most common in all answers («yes», «no», «2», «1», «white») -> to avoid learning question type based on answer
  - **Random**: correct answers from random questions

• Choices for Question : (Corrects+Plausibles+Populars)+ Randoms
Dataset(Answers)

- 23,234  3,770 unique one-word answer for real and abstract images
- Others are two or three words answers
- %38.37 and %40.66 of questions are yes/no for real and abstract images
- %58.83 and %55.86 of yes/no answers is «yes»
Dataset(Answers)

Agreement is
> %95 for «yes/no» questions
< %76 for other questions (due to synonyms, plurality etc)

Human agreement increases as confidence of answerer increases
Dataset (Are images necessary?)

- Is the image necessary?
  - What color is fire hydrant?
- For yes/no questions chance may saves you 😊
- For open-ended questions humans performance is %21
- >3 people thinks that %47.43 of questions can be answered
- >6 people thinks that %18.14 of questions can be answered
Dataset (Are images necessary?)
Dataset (Are images necessary?)
<table>
<thead>
<tr>
<th>Dataset (Are images necessary?)</th>
<th>Input</th>
<th>All</th>
<th>Yes/No</th>
<th>Number</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>Question</td>
<td>40.81</td>
<td>67.60</td>
<td>25.77</td>
<td>21.22</td>
</tr>
<tr>
<td></td>
<td>Question + Caption*</td>
<td>57.47</td>
<td>78.97</td>
<td>39.68</td>
<td>44.41</td>
</tr>
<tr>
<td></td>
<td>Question + Image</td>
<td>83.30</td>
<td>95.77</td>
<td>83.39</td>
<td>72.67</td>
</tr>
<tr>
<td>Abstract</td>
<td>Question</td>
<td>43.27</td>
<td>66.65</td>
<td>28.52</td>
<td>23.66</td>
</tr>
<tr>
<td></td>
<td>Question + Caption*</td>
<td>54.34</td>
<td>74.70</td>
<td>41.19</td>
<td>40.18</td>
</tr>
<tr>
<td></td>
<td>Question + Image</td>
<td>87.49</td>
<td>95.96</td>
<td>95.04</td>
<td>75.33</td>
</tr>
</tbody>
</table>

Captions are not enough for humans to answer questions about scene.

Kolmogorov-Smirnov Test

\[ p(\text{Captions vs (Q+A)}) < 0.001 \]
Model (Input Representation)

- **Image**
  - $I = \text{activations from last hidden layer of VGGNet as 4096-dim embedding}$
  - $\text{norm } I = l_2\text{normalized activations from last hidden layer of VGGNet}$

- **Question**
  - **BoW Q**
    - 1030 dim-embedding : Top 1000 popular words in questions + top 10 first, second, third words of questions
  - **LSTM Q:**
    - LSTM with one hidden layer
    - 1024 dim embedding for the question
    - 300 dim embedding for each question word
  - **Deeper LSTM Q:**
    - LSTM with two hidden layer
    - 2048 dim embedding for the question
    - with fc+tanh to transform 2048 dim to 1024 dim-embedding
Model (Generating Answer)

- **Multi Layer Perceptron:**
  - Bow Q + I
    - Concatenation
    - Bow Q + C
  - LSTM Q + I, LSTM Q + norm I
    - Fusion with point-wise multiplication
Model (Generating Answer)

Image Credit: Avi Singh
Model (Generating Answer)

LSTM: one hidden layer
output size 1024
each word size 300
Deeper LSTM: two hidden layer
output: 2048 > fc+tanh > 1024

MLP: 2 hidden layer fc network
1000 dropout(0.5) units tanh
end-to-end learning cross-entropy

Input Vocabulary: All question words
2-Channel VQA Model

Image Embedding

Convolution Layer + Non-Linearity

Pooling Layer

Convolution Layer + Non-Linearity

Pooling Layer

Fully-Connected MLP

4096-dim

Neural Network Softmax over top K answers

Question Embedding

“How many horses are in this image?”

1024-dim

Slide Credits: Agrawal
Ablation #1: Language-alone

Image Embedding

Convolution Layer + Non-Linearity
Pooling Layer
Convolution Layer + Non-Linearity
Pooling Layer

Fully-Connected MLP
1k output

Embedding Neural Network

Softmax over top K answers

Question Embedding

“How many horses are in this image?”

1024-dim

Slide Credits: Agrawal
Ablation #2: Vision-alone

Image Embedding

Neural Network
Softmax over top K answers

Question Embedding

“How many horses are in this”

Slide Credits: Agrawal
Open-Ended Task Accuracies

Human vs. Machine performance

Accuracy (in %)

Human: 83.30

Machine: 58.16

Room for improvement: 25.14
Results

<table>
<thead>
<tr>
<th></th>
<th>Open-Ended</th>
<th></th>
<th></th>
<th>Multiple-Choice</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Yes/No</td>
<td>Number</td>
<td>Other</td>
<td>All</td>
<td>Yes/No</td>
</tr>
<tr>
<td>BoW Q</td>
<td>48.09</td>
<td>75.66</td>
<td>36.70</td>
<td>27.14</td>
<td>53.68</td>
<td>75.71</td>
</tr>
<tr>
<td>I</td>
<td>28.13</td>
<td>64.01</td>
<td>00.42</td>
<td>03.77</td>
<td>30.53</td>
<td>69.87</td>
</tr>
<tr>
<td>BoW Q + I</td>
<td>52.64</td>
<td>75.55</td>
<td>33.67</td>
<td>37.37</td>
<td>58.97</td>
<td>75.59</td>
</tr>
<tr>
<td>LSTM Q</td>
<td>48.76</td>
<td>78.20</td>
<td>35.68</td>
<td>26.59</td>
<td>54.75</td>
<td>78.22</td>
</tr>
<tr>
<td>LSTM Q + I</td>
<td>53.74</td>
<td>78.94</td>
<td>35.24</td>
<td>36.42</td>
<td>57.17</td>
<td>78.95</td>
</tr>
<tr>
<td>deeper LSTM Q</td>
<td>50.39</td>
<td>78.41</td>
<td>34.68</td>
<td>30.03</td>
<td>55.88</td>
<td>78.45</td>
</tr>
<tr>
<td>deeper LSTM Q + norm I</td>
<td>57.75</td>
<td>80.50</td>
<td>36.77</td>
<td>43.08</td>
<td>62.70</td>
<td>80.52</td>
</tr>
<tr>
<td>Caption</td>
<td>26.70</td>
<td>65.50</td>
<td>02.03</td>
<td>03.86</td>
<td>28.29</td>
<td>69.79</td>
</tr>
<tr>
<td>BoW Q + C</td>
<td>54.70</td>
<td>75.82</td>
<td>40.12</td>
<td>42.56</td>
<td>59.85</td>
<td>75.89</td>
</tr>
</tbody>
</table>

**TABLE 2**: Accuracy of our methods for the open-ended and multiple-choice tasks on the VQA test-dev for real images. Q = Question, I = Image, C = Caption. (Caption and BoW Q + C results are on val).
Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources

Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, Anthony Dick
School of Computer Science, The University of Adelaide
{qi.wu01, p.wang, chunhua.shen, anton.vandenhengel, anthony.dick}@adelaide.edu.au

Compositional Memory for Visual Question Answering

Aiwen Jiang1,2  Fang Wang2  Fatih Porikli2  Yi Li*2,3
1ingxi Normal University  2NICTA and ANU  3Toyota Research Institute North America
aiwen.jiang@nicta.com.au  {fang.wang, fatih.porikli}@nicta.com.au  yilili@tema.toyota.com

Deep Compositional Question Answering with Neural Module Networks

Jacob Andreas  Marcus Rohrbach  Trevor Darrell  Dan Klein
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
{jda, rohrbach, trevor, klein}@cs.berkeley.edu

Stacked Attention Networks for Image Question Answering

Zichao Yang1, Xiaodong He2, Jianfeng Gao2, Li Deng2, Alex Smola1
1Carnegie Mellon University, 2Microsoft Research, Redmond, WA 98052, USA
cy@cs.cmu.edu, {xiaocheng, jfgao, deng}@microsoft.com, alex@smola.org

Simple Baseline for Visual Question Answering

Bolei Zhou1, Yuandong Tian2, Sainbayar Sukhbaatar2, Arthur Szlam2, and Rob Fergus2
1Massachusetts Institute of Technology  2Facebook AI Research

Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering

Haijuan Xu
UMass Lowell
hxu1@cs.uml.edu

Where To Look: Focus Regions for Visual Question Answering

Kevin J. Shih, Saurabh Singh, and Derek Hoiem
University of Illinois at Urbana-Champaign
{kjshih2, ssi, dhoiem}@illinois.edu

BC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering

Kan Chen
University of Southern California
kanchen@usc.edu

Wei Xu
Baidu Research - IDL
weixu@baidu.com

Haoyuan Gao
Baidu Research - IDL
haoyuangan@baidu.com

Jiang Wang
Baidu Research - IDL
jiangw@139.com

Liang-Chieh Chen
UCLA
lochen@cs.ucla.edu

Ram Nevatia
University of Southern California
nerav@usc.edu
## Current Leaderboard

<table>
<thead>
<tr>
<th>Team Name</th>
<th>Yes/No</th>
<th>Number</th>
<th>Other</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>snubi-naoverlabs[4]</td>
<td>82.23</td>
<td>38.22</td>
<td>46.99</td>
<td>60.6</td>
</tr>
<tr>
<td>MM_PaloAlto[3]</td>
<td>80.43</td>
<td>36.82</td>
<td>48.33</td>
<td>60.36</td>
</tr>
<tr>
<td>LV-NUS[2]</td>
<td>81.34</td>
<td>35.67</td>
<td>46.10</td>
<td>59.54</td>
</tr>
<tr>
<td>ACVT_Adeelaide[1]</td>
<td>81.07</td>
<td>37.12</td>
<td>45.83</td>
<td>59.44</td>
</tr>
<tr>
<td>vqateam-deeperLSTM_NormlizeCNN[6]</td>
<td>80.56</td>
<td>36.53</td>
<td>43.73</td>
<td>58.16</td>
</tr>
<tr>
<td>vqateam-lstm_cnn[7]</td>
<td>79.01</td>
<td>35.55</td>
<td>36.8</td>
<td>54.06</td>
</tr>
<tr>
<td>vqateam-q_lstm_alone[10]</td>
<td>78.12</td>
<td>34.94</td>
<td>26.99</td>
<td>48.89</td>
</tr>
<tr>
<td>vqateam-nearest_neighbor[8]</td>
<td>71.73</td>
<td>24.31</td>
<td>22</td>
<td>42.73</td>
</tr>
<tr>
<td>vqateam-prior_per_qtype[9]</td>
<td>71.17</td>
<td>35.63</td>
<td>9.32</td>
<td>37.55</td>
</tr>
<tr>
<td>vqateam-all_yes[5]</td>
<td>70.53</td>
<td>0.43</td>
<td>1.26</td>
<td>29.72</td>
</tr>
</tbody>
</table>
Questions & Discussion & Demo