Words, Pictures, and Common Sense

Devi Parikh
Georgia Tech
People coloring a street in rural Virginia.
It was a great event! It brought families out, and the whole community together.
Q. What are they coloring the street with?
A. Chalk
AI: What a nice picture! What event was this?

User: “Color College Avenue”. It was a lot of fun!

AI: I am sure it was! Do they do this every year?

User: I wish they would. I don’t think they’ve organized it again since 2012.

...
Words

Pictures

Common Sense
Applications

Pictures are everywhere
Words are how we communicate
Applications

Interact with, organize, and navigate visual data
Applications

Leverage multi-modal information on the web
Applications

Aid visually-impaired users
Applications

Aid visually-impaired users

FACEBOOK’S AI CAN CAPTION PHOTOS FOR THE BLIND ON ITS OWN
Applications

Peter just uploaded a picture from his vacation in Hawaii

Great, is he at the beach?

No, on a mountain
Applications
Summarize visual data for analysts

Did anyone enter this room last week?

Yes, 127 instances logged on camera

Were any of them carrying a black bag?

...
Applications

Natural language instructions to an agent

Is there smoke in any room around you?

Yes, in one room

Go there and look for people

...
Words

Visually

Grounded

Dialog

Common Sense

Pictures
Visual Question Answering (VQA)
Visual Question Answering (VQA)

What is the mustache made of?
Visual Question Answering (VQA)

What is the mustache made of?

AI System
Visual Question Answering (VQA)

What is the mustache made of?

AI System

bananas
Visual Question Answering (VQA)

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?
>0.25 million images
254,721 images (MS COCO)
50,000 scenes
>0.25 million images

>0.76 million questions
Questions

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

We have built a system that can recognize the scene (e.g., kitchen, beach) and can answer questions about it such as: “What is the man doing?” “Is the man eating?” “What is the man holding?” We feel confident that even a smart robot should not be able to answer these questions correctly, unless a human is present to guide it.

Important: The human should remain seated and should not be able to intervene in the interaction. The questions asked by the human should be natural and difficult for the robot to answer.

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.
>0.25 million images

>0.76 million questions

~10 million answers

>20 person-job-years
Taxing the Turkers

• Beware also the lasting effects of doing too many --for hours after the fact you will not be able to look at any photo without automatically generating a mundane question for it.

• If I were in possession of state secrets they could be immediately tortured out of me with the threat of being shown images of: skateboards, trains, Indian food and [long string of expletives] giraffes.

• (Please...I will tell you everything...just no more giraffes...)
Answers

[Bar chart with various categories and answers]

Slide credit: Devi Parikh
Human Accuracy, Inter-Human Agreement

Human agreement: 83%

First model (Summer 2015): 54%

State-of-the-art machine accuracy: 68%
Visual Question Answering (VQA)

Ask any question about this image

www.visualqa.org
Papers using VQA

Ask Me Anything: Free-form Visual Question Answering
Based on Knowledge from External Sources

Simple Baseline for Visual Question Answering

Academia, industry, start ups
Dataset, Code

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

Kan Chen
University of Southern California
kanchen@usc.edu

Haoyuan Gao
Baidu Research - IDL
gaohaoyuan@baidu.com

Wei Xu
Baidu Research - IDL
wei.xu@baidu.com

Jiang Wang
Baidu Research - IDL
wangjiang03@baidu.com

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

Ram Nevatia
University of Southern California
nevatia@usc.edu

Stacked Attention Networks for Image Question Answering

Zichao Yang\textsuperscript{1}, Xiaodong He\textsuperscript{2}, Jianfeng Gao\textsuperscript{2}, Li Deng\textsuperscript{2}, Alex Smola\textsuperscript{1}
\textsuperscript{1}Carnegie Mellon University, \textsuperscript{2}Microsoft Research, Redmond, WA 98052, USA
py@cs.cmu.edu, \{xiaohe, jfgao, deng\}@microsoft.com, alex@smola.org

Slide credit: Devi Parikh
Recent progress in computer vision and natural language processing has demonstrated that lower-level tasks are much closer to being solved. We believe that the time is ripe to pursue higher-level tasks, one of which is Visual Question Answering (VQA). Visual question answering aims to understand the content of an image and answer questions about it.
### VQA Challenge @ CVPR16

<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>UC Berkeley &amp; Sony[1,2]</td>
<td>83.24</td>
<td>39.47</td>
<td>58</td>
<td>68.47</td>
</tr>
<tr>
<td>Naver Labs[10]</td>
<td>83.31</td>
<td>38.7</td>
<td>54.62</td>
<td>64.79</td>
</tr>
<tr>
<td>DLAI[3]</td>
<td>83.25</td>
<td>40.07</td>
<td>52.09</td>
<td>63.68</td>
</tr>
<tr>
<td>snubi-saverlabs[22]</td>
<td>83.16</td>
<td>39.14</td>
<td>51.33</td>
<td>63.18</td>
</tr>
<tr>
<td>POSTECH[11]</td>
<td>81.67</td>
<td>38.16</td>
<td>52.79</td>
<td>63.17</td>
</tr>
<tr>
<td>Brandeis[11]</td>
<td>82.11</td>
<td>37.73</td>
<td>51.91</td>
<td>62.68</td>
</tr>
<tr>
<td>VTComputer/Vision[39]</td>
<td>78.95</td>
<td>38.22</td>
<td>51.95</td>
<td>62.06</td>
</tr>
<tr>
<td>MIL-U[17]</td>
<td>81.98</td>
<td>37.56</td>
<td>49.76</td>
<td>61.77</td>
</tr>
<tr>
<td>iLab[22]</td>
<td>81.53</td>
<td>39.27</td>
<td>49.81</td>
<td>61.69</td>
</tr>
<tr>
<td>SHB_1028[13]</td>
<td>82.07</td>
<td>36.81</td>
<td>47.77</td>
<td>60.76</td>
</tr>
<tr>
<td>MMCX[8]</td>
<td>80.43</td>
<td>36.62</td>
<td>48.33</td>
<td>60.36</td>
</tr>
<tr>
<td>VT_GV_Jlasen[29]</td>
<td>80.56</td>
<td>38.14</td>
<td>47.87</td>
<td>60.33</td>
</tr>
<tr>
<td>LV-NUS[30]</td>
<td>81.34</td>
<td>35.67</td>
<td>46.1</td>
<td>59.54</td>
</tr>
<tr>
<td>ACVT_Adeleide[1]</td>
<td>81.07</td>
<td>37.12</td>
<td>45.83</td>
<td>59.44</td>
</tr>
<tr>
<td>UC Berkeley (CINN)[19]</td>
<td>80.98</td>
<td>37.48</td>
<td>45.81</td>
<td>59.44</td>
</tr>
<tr>
<td>CNNAS[4]</td>
<td>81.04</td>
<td>36.44</td>
<td>45.76</td>
<td>59.33</td>
</tr>
<tr>
<td>sen[24]</td>
<td>79.11</td>
<td>36.41</td>
<td>46.42</td>
<td>58.85</td>
</tr>
<tr>
<td>UC Berkeley (NIN)[19]</td>
<td>81.18</td>
<td>37.7</td>
<td>44.01</td>
<td>58.66</td>
</tr>
<tr>
<td>globai_vision[22]</td>
<td>78.24</td>
<td>36.27</td>
<td>46.32</td>
<td>58.43</td>
</tr>
<tr>
<td>vqatm-deeper_LSTM_NormalizeCNN[21]</td>
<td>80.56</td>
<td>36.53</td>
<td>43.73</td>
<td>58.16</td>
</tr>
<tr>
<td>Mujibra hassan[8]</td>
<td>80.28</td>
<td>36.92</td>
<td>42.24</td>
<td>57.36</td>
</tr>
<tr>
<td>RIT[12]</td>
<td>78.82</td>
<td>35.97</td>
<td>42.13</td>
<td>56.61</td>
</tr>
<tr>
<td>Boie[2]</td>
<td>76.76</td>
<td>34.98</td>
<td>42.62</td>
<td>55.69</td>
</tr>
<tr>
<td>UPV_UBI[14]</td>
<td>78.88</td>
<td>36.33</td>
<td>40.27</td>
<td>55.77</td>
</tr>
<tr>
<td>arP[21]</td>
<td>78.1</td>
<td>35.3</td>
<td>40.27</td>
<td>55.34</td>
</tr>
<tr>
<td>vqatm-klm_cnn[25]</td>
<td>79.01</td>
<td>35.55</td>
<td>36.8</td>
<td>54.06</td>
</tr>
<tr>
<td>UPC[17]</td>
<td>78.05</td>
<td>35.53</td>
<td>36.7</td>
<td>53.62</td>
</tr>
<tr>
<td>vqatm-nearest_neighbor[20]</td>
<td>71.73</td>
<td>24.31</td>
<td>22</td>
<td>42.73</td>
</tr>
<tr>
<td>vqatm-prior_per_qtype[20]</td>
<td>71.17</td>
<td>35.63</td>
<td>9.32</td>
<td>37.55</td>
</tr>
<tr>
<td>vqatm-all-yes[23]</td>
<td>70.53</td>
<td>0.43</td>
<td>1.26</td>
<td>29.72</td>
</tr>
</tbody>
</table>

~30 teams
What such a model can’t do

How many vegetarian slices are left in the pizza box?
It can’t count...

How many vegetarian slices are left in the pizza box?
It doesn’t have commonsense / knowledge...

How many vegetarian slices are left in the pizza box?
It can’t reason...

How many vegetarian slices are left in the pizza box?
It doesn’t leverage compositionality...

How many vegetarian slices are left in the pizza box?
It lacks consistency...

How many vegetarian slices are left in the pizza box?
Visual Dialog
VisDial Dataset

Live Two-Person Chat on Amazon Mechanical Turk

Questioner  Answerer
VisDial Dataset
Live Two-Person Chat on Amazon Mechanical Turk

Caption: The man is riding his bicycle on the sidewalk
You have to ASK Questions about the image.

Caption: The man is riding his bicycle on the sidewalk
You have to ANSWER questions about the image.
VisDial Dataset

- ~100k images
- 1 dialog per image
- 2 people
- 10 rounds per person
Qualitative Results

Hi, I am a Visual Chatbot, capable of answering a sequence of questions about images. Please upload an image and fire away!

Drag and Drop Image here

Start typing question here...
Applications

Peter just uploaded a picture from his vacation in Hawaii

Great, is he at the beach?

No, on a mountain

Image may contain: two people, smiling, sunglasses, sky, outdoor, water

Slide credit: Devi Parikh
Looking forward: Visual Dialog for Action
Applications
Natural language instructions to an agent

Is there smoke in any room around you?

Yes, in one room

Go there and look for people

...
Create a children's illustration!

Please help create our illustration by completing the scenes below. Use your imagination! Clipart may be added by dragging the clipart to the scene, and removed by dragging it off. The clipart may be resized when flipped and rotated clipart may only be added once. Please use at least 6 pieces of clipart in each scene. You will be asked to complete 3 different scenes. Press "Next" when finished with the current scene and "Done" when all are finished. Thanks!
Describer sees a reference image

Actor does now
Actor: What is going on in the image overall?

Describer: Two people are playing

Actor: Is Mike to Jenny's left?

Describer: No, on the right
Actor: What is going on in the image overall?

Describer: Two people are playing

Actor: Is Mike to Jenny's left?

Describer: No, on the right
Describer: There is an airplane
Actor: What is going on in the image overall?

Describer: Two people are playing

Actor: Is Mike to Jenny's left?

Describer: No, on the right
Describer: There is an airplane

Actor: What else is there?

Describer: Sun is on the top right corner
Describer: And a tree on the right side with fruits

Actor: Is the plane going towards the sun?

Describer: Yes
Actor: What is going on in the image overall?

Describer: Two people

Actor: Is Mike to Jenny's left?

Describer: No, on the right
Describer: There is an airplane

Actor: What else is there?

Describer: Sun is on the top right corner
Describer: And a tree on the right side with fruits

Actor: Is the plane going towards the sun?

Describer: Yes
Describer: Jenny is standing
Describer: They are both standing in a sandbox

Actor: What expressions do they have?

Describer: Jenny is smiling while Mike is surprised

Actor: So there are just 6 objects in the image?

Describer: and there is an owl sitting on the tree
Describer: 7 objects
Describer: The plane is farther away from the sun. The whole tree is visible
Common Sense
Man in blue wetsuit is surfing on wave
Karpathy and Fei-Fei (Stanford) 2015

A group of young people playing a game of Frisbee
Vinyals et al. (Google) 2015

A car is parked in the middle of nowhere
Kiros et al. (University of Toronto) 2015

A pot of broccoli on a stove.
Fang et al. (Microsoft Research) 2015
A man is rescued from his truck that is hanging dangerously from a bridge.
A man is *rescued* from his truck that is hanging *dangerously* from a bridge.
Learning Common Sense

• Text
  – Reporting bias
## Reporting bias in text

<table>
<thead>
<tr>
<th>Word</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>spoke</td>
<td>11,577,917</td>
<td>244,458</td>
</tr>
<tr>
<td>laughed</td>
<td>3,904,519</td>
<td>169,347</td>
</tr>
<tr>
<td>murdered</td>
<td>2,843,529</td>
<td>11,284</td>
</tr>
<tr>
<td>inhaled</td>
<td>984,613</td>
<td>4,412</td>
</tr>
<tr>
<td>breathed</td>
<td>725,034</td>
<td>34,912</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>hugged</td>
<td>610,040</td>
<td>10,378</td>
</tr>
<tr>
<td>blinked</td>
<td>390,692</td>
<td>20,624</td>
</tr>
<tr>
<td>was late</td>
<td>368,922</td>
<td>31,168</td>
</tr>
<tr>
<td>exhaled</td>
<td>168,985</td>
<td>3,490</td>
</tr>
<tr>
<td>was punctual</td>
<td>5,045</td>
<td>511</td>
</tr>
</tbody>
</table>

[Gordon et al. 2013]
Reporting bias in text

<table>
<thead>
<tr>
<th>Word</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>spoke</td>
<td>11,577,917</td>
<td>244,458</td>
</tr>
<tr>
<td>laughed</td>
<td>3,904,519</td>
<td></td>
</tr>
<tr>
<td>murdered</td>
<td>2,843,529</td>
<td>11,284</td>
</tr>
<tr>
<td>inhaled</td>
<td>984,613</td>
<td>4,412</td>
</tr>
<tr>
<td>breathed</td>
<td>725,034</td>
<td>34,912</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>hugged</td>
<td>610,040</td>
<td>10,378</td>
</tr>
<tr>
<td>exhaled</td>
<td>368,922</td>
<td>31,168</td>
</tr>
<tr>
<td>was late</td>
<td>390,692</td>
<td>20,624</td>
</tr>
<tr>
<td>was punctual</td>
<td>5,045</td>
<td>511</td>
</tr>
</tbody>
</table>

inhale:exhale = 6:1

[Gordon et al. 2013]
### Reporting bias in text

<table>
<thead>
<tr>
<th>Word</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>spoke</td>
<td>11,577,917</td>
<td>244,458</td>
</tr>
<tr>
<td>laughed</td>
<td>3,904,519</td>
<td>169,347</td>
</tr>
<tr>
<td><strong>murdered</strong></td>
<td><strong>2,843,529</strong></td>
<td><strong>11,284</strong></td>
</tr>
<tr>
<td>inhaled</td>
<td>984,613</td>
<td>4,412</td>
</tr>
<tr>
<td>breathed</td>
<td>725,034</td>
<td>34,912</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>hugged</td>
<td>610,040</td>
<td>10,378</td>
</tr>
<tr>
<td>blinked</td>
<td>390,692</td>
<td>20,624</td>
</tr>
<tr>
<td>was late</td>
<td>368,922</td>
<td>31,168</td>
</tr>
<tr>
<td><strong>exhaled</strong></td>
<td><strong>168,985</strong></td>
<td></td>
</tr>
<tr>
<td>was punctual</td>
<td>5,045</td>
<td>511</td>
</tr>
</tbody>
</table>

\[
murder:exhale = 17:1
\]

[Gordon et al. 2013]
## Reporting bias in text

<table>
<thead>
<tr>
<th>Body Part</th>
<th>Teraword</th>
<th>Knext</th>
<th>Body Part</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head</td>
<td>18,907,427</td>
<td>1,332,154</td>
<td>Liver</td>
<td>246,937</td>
<td>10,474</td>
</tr>
<tr>
<td>Eye(s)</td>
<td>18,455,030</td>
<td>1,090,640</td>
<td>Kidney(s)</td>
<td>183,973</td>
<td>5,014</td>
</tr>
<tr>
<td>Arm(s)</td>
<td>6,345,039</td>
<td>458,018</td>
<td>Spleen</td>
<td>47,216</td>
<td>1,414</td>
</tr>
<tr>
<td>Ear(s)</td>
<td>3,543,711</td>
<td>230,367</td>
<td>Pancreas</td>
<td>24,230</td>
<td>1,140</td>
</tr>
<tr>
<td>Brain</td>
<td>3,277,326</td>
<td>260,863</td>
<td>Gallbladder</td>
<td>17,419</td>
<td>1,556</td>
</tr>
</tbody>
</table>

[Gordon et al. 2013]
Reporting bias in text

<table>
<thead>
<tr>
<th>Body Part</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head</td>
<td>18,907,427</td>
<td>1,332,154</td>
</tr>
<tr>
<td>Eye(s)</td>
<td>18,455,030</td>
<td>1,090,640</td>
</tr>
<tr>
<td>Arm(s)</td>
<td>3,543,711</td>
<td>230,367</td>
</tr>
<tr>
<td>Ear(s)</td>
<td>3,277,326</td>
<td>260,863</td>
</tr>
<tr>
<td>Brain</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Body Part</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver</td>
<td>246,937</td>
<td>10,474</td>
</tr>
<tr>
<td>Kidney(s)</td>
<td>183,973</td>
<td>5,014</td>
</tr>
<tr>
<td>Pancreas</td>
<td>24,230</td>
<td>1,140</td>
</tr>
<tr>
<td>Gallbladder</td>
<td>17,419</td>
<td>1,556</td>
</tr>
</tbody>
</table>

People have heads:gallbladders = 1085:1

[Gordon et al. 2013]
Learning Common Sense

• Text
  – Reporting bias

• From structure in our visual world?
Two professors converse in front of a blackboard.
Two professors stand in front of a blackboard.
Two professors converse in front of a blackboard.
Challenges

• Lacking visual density
• Annotations are expensive
• Computer vision doesn’t work well enough
Is photorealism necessary?
Jenny

Mike
Mike fights off a bear by giving him a hotdog while Jenny runs away.
Commonsense Tasks

• Text-based tasks
Fill-in-the-blank:

Mike is having lunch when he sees a bear.

A. Mike orders a pizza.
B. Mike hugs the bear.
C. Bears are mammals.
D. Mike tries to hide.
Key idea

• Imagine the scene behind the text
• Reason about the visual interpretation of the text, not just the text alone
Visual Paraphrasing: Are these two descriptions describing the same scene?

1. Jenny was going to throw her pie at Mike.

2. Jenny is very angry. Jenny is holding a pie.

[Lin and Parikh, CVPR 2015]
Approach: Imagination

Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.

A. There is a tree near a table.
B. The brown dog is standing next to Mike.
C. The sun is in the sky.
D. Jenny is standing dangerously on the swing.
There is a tree near a table.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.

A. There is a tree near a table.
B. The brown dog is standing next to Mike.
C. The sun is in the sky.
D. Jenny is standing dangerously on the swing.
The brown dog is standing next to Mike.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.

A. There is a tree near a table.
B. The brown dog is standing next to Mike.
C. The sun is in the sky.
D. Jenny is standing dangerously on the swing.
A. There is a tree near a table.
B. The brown dog is standing next to Mike.
C. The sun is in the sky.
D. Jenny is standing dangerously on the swing.
Approach: Imagination

Jenny is standing dangerously on the swing.
Mike is wearing a blue cap.
Mike is telling Jenny to get off the swing.

A. There is a tree near a table.
B. The brown dog is standing next to Mike.
C. The sun is in the sky.
D. Jenny is standing dangerously on the swing.

Imagined scenes need not be photorealistic but rich in semantics.
A. There is a tree near a table.
B. The brown dog is standing next to Mike.
C. The sun is in the sky.
D. Jenny is standing dangerously on the swing.
Approach: Joint Text + Visual Reasoning

Jenny is standing dangerously on the swing. Mike is wearing a blue cap. Mike is telling Jenny to get off the swing.

There is a tree near a table. Mike is wearing a blue cap. Mike is telling Jenny to get off the swing.
“This terrified woman's home is being invaded by mice as the cat sleeps.”
“The man is about to trip on his child's car and spill wine on his wife.”
Summary: Visual Dialog

• Applications
  – Today’s chatbots are blind!

• AI
  – Vision
  – Language
  – Attention
  – Reasoning
  – External knowledge
  – Common Sense
  – Action, Manipulation
Thank you.