Visual Question and Answering
Lecture 8

Slides from Devi Parikh, Dhruv Bhatra, Ethan Perez, Jacob Andreas, Marcus Rorbach & others
Visual Question Answering

What is the mustache made of?

Devi Parikh
Virginia Tech
Visual Question Answering (VQA)

Task
• Given
  – An image
  – A natural language open-ended question
• Generate
  – A natural language answer
Visual Question Answering (VQA)

Ask any question about this image

www.visualqa.org
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?
Visual Question Answering (VQA)

- Details of the image
- Common sense + knowledge base
- Task-driven
- Holy-grail of semantic image understanding
“Can machines think”

Q: Please write me a sonnet on the subject of the Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764.
A: (Pause about 30 seconds and then give as answer) 105621.
Q: How many slices of pizza are there?  
A: 6
Visual Turing Test [Geman 2014]

• 2591 street city images
Vocabulary
- Types of objects
  - People, vehicles, building, windows, doors
- Type-dependent attributes
  - Clothing and activities of people
  - Types and colors of vehicles
- Type-dependent relationships
  - Ordered: person entering a building
  - Unordered: two people walking together

Questions
- Existence
- Uniqueness
- Attribute
- Relationship

Story line
Query generator
Human-in-the-loop

No NLP required, vision is key
DAQUAR [Malinowski 2014]

• DAtaset for QUestion Answering on Real-world images (DAQUAR)
• 1449 images from NYU v2
DAQUAR [Malinowski 2014]

- Synthetic QA pairs
  - 140 training
  - 280 test

<table>
<thead>
<tr>
<th>Description</th>
<th>Template</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>counting</td>
<td>How many {object} are in {image_id}?</td>
<td>How many cabinets are in image1?</td>
</tr>
<tr>
<td>counting and colors</td>
<td>How many {color} {object} are in {image_id}?</td>
<td>How many gray cabinets are in image1?</td>
</tr>
<tr>
<td>room type</td>
<td>Which type of the room is depicted in {image_id}?</td>
<td>Which type of the room is depicted in image1?</td>
</tr>
<tr>
<td>superlatives</td>
<td>What is the largest {object} in {image_id}?</td>
<td>What is the largest object in image1?</td>
</tr>
<tr>
<td>counting and colors</td>
<td>How many {color} {object}?</td>
<td>How many black bags?</td>
</tr>
<tr>
<td>negations type 1</td>
<td>Which images do not have {object}?</td>
<td>Which images do not have sofa?</td>
</tr>
<tr>
<td>negations type 2</td>
<td>Which images are not {room_type}?</td>
<td>Which images are not bedroom?</td>
</tr>
<tr>
<td>negations type 3</td>
<td>Which images have {object} but do not have a {object}?</td>
<td>Which images have desk but do not have a lamp?</td>
</tr>
</tbody>
</table>
• Human QA pairs
  – 6794 training
  – 5675 test
• Valid answers
  – Colors, numbers, objects, or sets
**DAQUAR [Malinowski 2014]**

**QA: (What is behind the table?, window)**
Spatial relation like ‘behind’ are dependent on the reference frame. Here the annotator uses observer-centric view.

**QA: (what is beneath the candle holder, decorative plate)**
Some annotators use variations on spatial relations that are similar, e.g. ‘beneath’ is closely related to ‘below’.

**QA: (what is in front of the wall divider?, cabinet)**
Annotators use additional properties to clarify object references (i.e. wall divider). Moreover, the perspective plays an important role in these spatial relations interpretations.

**QA: (what is behind the table?, sofa)**
Spatial relations exhibit different reference frames. Some annotations use observer-centric, others object-centric view.

The annotators are using different names to call the same things. The names of the brown object near the bed include ‘night stand’, ‘stool’, and ‘cabinet’.

QA1: (How many doors are in the image?, 1)
QA2: (How many doors are in the image?, 5)
Different interpretation of ‘door’ results in different counts: 1 door at the end of the hall vs. 5 doors including lockers.
DAQUAR [Malinowski 2014]

- Accuracy
- Wu-Palmer similarity (WUPS)
  - WUPS 0.0
  - WUPS 0.9
Toronto COCO-QA [Ren 2015]

• COCO dataset
• Caption $\rightarrow$ QA pair (automatically)
  – 123287 images
  – 78736 train questions
  – 38948 test questions
• 4 types of questions:
  – object, number, color, location
• Answers are all one-word
Toronto COCO-QA [Ren 2015]

COCOQA 5078
How many leftover donuts is the red bicycle holding?
Ground truth: three

COCOQA 1238
What is the color of the tee-shirt?
Ground truth: blue

COCOQA 26088
Where is the gray cat sitting?
Ground truth: window
Q. The very old looking what is on display?
A. pot

Q. What swim in the ocean near two large ferries?
A. ducks

Q. What next to the large umbrella attached to a table?
A. trees
Toronto COCO-QA [Ren 2015]

- Accuracy
- Wu-Palmer similarity (WUPS)
  - WUPS 0.0
  - WUPS 0.9
>0.25 million images
50,000 scenes
>0.25 million images

>0.76 million questions

~10 million answers
Questions

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

We have built
kitchen, beach

Ask a question
IMPORTANT: Think the question will

- 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.
>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

• 38% of questions are binary yes/no

• 99% questions have answers <= 3 words
  – Evaluation is feasible
  – 23k unique 1 word answers
Evaluation Formats

• Open answer
  – Input = image, question

• Multiple choice
  – Input = image, question, 18 answer options
  – Avoids language generation
  – Evaluation (even more) feasible
  – Options = \{correct, plausible, popular, random\} answers
Q. What is he playing?

(a) guitar
(b) drums
(c) baseball
**Accuracy Metric**

\[
\text{Acc}(\text{ans}) = \min \left\{ \frac{\text{#humans that said } \text{ans}}{3}, 1 \right\}
\]

### Open-Ended/Multiple-Choice/Ground-Truth

**Q: WHAT OBJECT IS THIS**

<table>
<thead>
<tr>
<th>Ground Truth Answers:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) television</td>
<td>(6) television</td>
</tr>
<tr>
<td>(2) tv</td>
<td>(7) television</td>
</tr>
<tr>
<td>(3) tv</td>
<td>(8) tv</td>
</tr>
<tr>
<td>(4) tv</td>
<td>(9) tv</td>
</tr>
<tr>
<td>(5) television</td>
<td>(10) television</td>
</tr>
</tbody>
</table>

**Q: How old is this TV?**

<table>
<thead>
<tr>
<th>Ground Truth Answers:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) 20 years</td>
<td>(6) old</td>
</tr>
<tr>
<td>(2) 35</td>
<td>(7) 80 s</td>
</tr>
<tr>
<td>(3) old</td>
<td>(8) 30 years</td>
</tr>
<tr>
<td>(4) more than thirty years</td>
<td>(9) 15 years</td>
</tr>
<tr>
<td>old</td>
<td>(10) very old</td>
</tr>
</tbody>
</table>

**Q: Is this TV upside-down?**

<table>
<thead>
<tr>
<th>Ground Truth Answers:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) yes</td>
<td>(6) yes</td>
</tr>
<tr>
<td>(2) yes</td>
<td>(7) yes</td>
</tr>
<tr>
<td>(3) yes</td>
<td>(8) yes</td>
</tr>
<tr>
<td>(4) yes</td>
<td>(9) yes</td>
</tr>
<tr>
<td>(5) yes</td>
<td>(10) yes</td>
</tr>
</tbody>
</table>
Human Accuracy, Inter-Human Agreement

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Input</th>
<th>All</th>
<th>Yes/No</th>
<th>Number</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Real</strong></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><strong>Abstract</strong></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>
## Human Accuracy, Inter-Human Agreement

<table>
<thead>
<tr>
<th>Dataset</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>
VQA Common Sense

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.

To answer this question, is commonsense required?
- 1. yes
- 2. no
VQA Common Sense

• Our best algorithm has* 17% common sense!
• Average common sense required = 31%.

* as estimated by untrained crowd-sourced workers in uncontrolled environment
How Old Do You Think a Person Needs to be to Answer These Questions?

We will present you with a series of questions about images. For each question, please select the youngest age group that you think a person must be in order to be able to correctly answer the question.

To answer this question, I would expect a person to have to at least be a:

- 1. toddler (3-4)
- 2. younger child (5-8)
- 3. older child (9-12)
- 4. teenager (13-17)
- 5. adult (18+)
<table>
<thead>
<tr>
<th>3-4 (15.3%)</th>
<th>5-8 (39.7%)</th>
<th>9-12 (28.4%)</th>
<th>13-17 (11.2%)</th>
<th>18+ (5.5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is that a bird in the sky?</td>
<td>How many pizzas are shown?</td>
<td>Where was this picture taken?</td>
<td>Is he likely to get mugged if he walked down a dark alleyway like this?</td>
<td>What type of architecture is this?</td>
</tr>
<tr>
<td>What color is the shoe?</td>
<td>What are the sheep eating?</td>
<td>What ceremony does the cake commemorate?</td>
<td>Is this a vegetarian meal?</td>
<td>Is this a Flemish bricklaying pattern?</td>
</tr>
<tr>
<td>How many zebras are there?</td>
<td>What color is his hair?</td>
<td>Are these boats too tall to fit under the bridge?</td>
<td>What type of beverage is in the glass?</td>
<td>How many calories are in this pizza?</td>
</tr>
<tr>
<td>Is there food on the table?</td>
<td>What sport is being played?</td>
<td>What is the name of the white shape under the batter?</td>
<td>Can you name the performer in the purple costume?</td>
<td>What government document is needed to partake in this activity?</td>
</tr>
<tr>
<td>Is this man wearing shoes?</td>
<td>Name one ingredient in the skillet.</td>
<td>Is this at the stadium?</td>
<td>Besides these humans, what other animals eat here?</td>
<td>What is the make and model of this vehicle?</td>
</tr>
<tr>
<td>Question</td>
<td>Average Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what brand</td>
<td>12.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>why</td>
<td>11.18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what type</td>
<td>11.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what kind</td>
<td>10.55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is this</td>
<td>10.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what does</td>
<td>10.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what time</td>
<td>9.81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>who</td>
<td>9.58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>where</td>
<td>9.54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>which</td>
<td>9.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>does</td>
<td>9.29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>do</td>
<td>9.23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what is</td>
<td>9.11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what are</td>
<td>9.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>are</td>
<td>8.65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is the</td>
<td>8.52</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is there</td>
<td>8.24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what sport</td>
<td>8.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>how many</td>
<td>7.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what animal</td>
<td>6.74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what color</td>
<td>6.6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
VQA Age

- Our best algorithm =* 4.84 years old!
- Average “age of questions” = 8.98 years.

* age as estimated by untrained crowd-sourced workers in uncontrolled environment
Datasets

Models

Current Status

Ongoing Efforts
Challenges in VQA

- Image representation
- Language representation
- Combining the modalities
- Attention
- Question-specific reasoning
- External knowledge
Basic Approach

Figure from de Vries et al. “Modulating early visual processing by language.” arXiv 2017.

Slide credit: Devi Parikh
Input: Image, Question
Output: Answer

Image:
- Convolutional Neural Network (CNN)
  [Fukushima 1980, LeCun et al. 1989]

Question:
- Recurrent Neural Network
- Specifically, a Long Short-Term Memory (LSTM)
  [Hochreiter & Schmidhuber, 1997]

Output: 1 of K most common answers
"How many horses are in this image?"
Ablation #1: Language-alone

“How many horses are in this image?”

2x2x512 LSTM

“2”

Point-wise multiplication

Fully-Connected

Softmax over 1000 most frequent answers

4096 output units from last hidden layer (VGG-19, Normalized)

1024

1024

1000

1000

Fully-Connected

Ablation #1: Language-alone
## Results

<table>
<thead>
<tr>
<th></th>
<th>Open-Answer</th>
<th></th>
<th></th>
<th></th>
<th>Multiple-Choice</th>
<th></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>
<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>
<td>53.68</td>
<td>75.71</td>
<td>37.05</td>
<td>38.64</td>
</tr>
<tr>
<td>Image</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>
<td>00.45</td>
<td>03.76</td>
</tr>
<tr>
<td>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>
<td>34.35</td>
<td>50.33</td>
</tr>
<tr>
<td>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>
<td>35.91</td>
<td>41.13</td>
</tr>
<tr>
<td>LSTM Q+I</td>
<td><strong>57.75</strong></td>
<td><strong>80.5</strong></td>
<td><strong>36.77</strong></td>
<td><strong>43.08</strong></td>
<td><strong>62.7</strong></td>
<td><strong>80.52</strong></td>
<td><strong>38.22</strong></td>
<td><strong>53.01</strong></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>
<td>02.06</td>
<td>03.82</td>
</tr>
<tr>
<td>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>
<td>41.16</td>
<td>52.53</td>
</tr>
</tbody>
</table>

- “yes” 29.27
- k-NN 40.61

- Multiple-Choice > Open-Ended
- Question alone does quite well
  - Better than humans
- Image helps

Code available!
## Results

<table>
<thead>
<tr>
<th></th>
<th>Open-Ended</th>
<th></th>
<th></th>
<th></th>
<th>Multiple-Choice</th>
<th></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>
<td>Number</td>
<td>Other</td>
</tr>
<tr>
<td>prior (&quot;yes&quot;)</td>
<td>29.66</td>
<td>70.81</td>
<td>00.39</td>
<td>01.15</td>
<td>29.66</td>
<td>70.81</td>
<td>00.39</td>
<td>01.15</td>
</tr>
<tr>
<td>per Q-type prior</td>
<td>37.54</td>
<td>71.03</td>
<td>35.77</td>
<td>09.38</td>
<td>39.45</td>
<td>71.02</td>
<td>35.86</td>
<td>13.34</td>
</tr>
<tr>
<td>nearest neighbor</td>
<td>42.70</td>
<td>71.89</td>
<td>24.36</td>
<td>21.94</td>
<td>48.49</td>
<td>71.94</td>
<td>26.00</td>
<td>33.56</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>
<td>37.05</td>
<td>38.64</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>
<td>00.45</td>
<td>03.76</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>
<td>34.35</td>
<td>50.33</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>
<td>36.82</td>
<td>38.78</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>
<td>35.80</td>
<td>43.41</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>
<td>35.91</td>
<td>41.13</td>
</tr>
<tr>
<td>deeper LSTM Q + norm I</td>
<td><strong>57.75</strong></td>
<td><strong>80.50</strong></td>
<td><strong>36.77</strong></td>
<td><strong>43.08</strong></td>
<td><strong>62.70</strong></td>
<td><strong>80.52</strong></td>
<td><strong>38.22</strong></td>
<td><strong>53.01</strong></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>
<td>02.06</td>
<td>03.82</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>
<td>41.16</td>
<td>52.53</td>
</tr>
</tbody>
</table>
Demo

Ask any question about this image

www.visualqa.org
[Malinowski 2015]

Slide credit: Devi Parikh
What is the cat doing? 

Embedding

LSTM

<BOA> Sitting on the umbrella <EOA>

CNN

Fusing

Intermediate

Softmax

Slide credit: Devi Parikh

[Gao 2015]
Challenges in VQA

• Image representation
• Language representation
• Combining the modalities
• Attention
• Question-specific reasoning
Is this going to be a feast?

CNN

Is?
feast
going to be
...

LSTM

spoon
plate
bowl
food
corn
...

person

☐ All elements can interact
☐ Multiplicative interaction

How to Combine
Image Representation and Question Representation?
How to Combine Image Representation and Question Representation?

Is this going to be a feast?

- CNN
- LSTM
- All elements can interact
- Multiplicative interaction
  - Difficult to learn output classification

Yes
Is this going to be a feast?

All elements can interact
- Multiplicative interaction
- Difficult to learn input embedding
How to Combine Image Representation and Question Representation?

Is this going to be a feast?

Outer Product / Bilinear Pooling [Lin ICCV 2015]

Yes

☑ All elements can interact
☑ Multiplicative interaction

Bilinear CNN models for fine-grained visual recognition. ICCV 2015

Slide credit: Akira Fukui and Marcus Rohrbach
How to Combine Image Representation and Question Representation?

Is this going to be a feast?

Outer Product / Bilinear Pooling

- CNN
- spoon
- plate
- bowl
- table
- food
- corn
- person

Is?
- feast
- going to be
- ...

2048

4 million

4 million x 1000

- Yes

☑ All elements can interact
☑ Multiplicative interaction
☐ High #activations & computation
☐ High #parameters


Slide credit: Akira Fukui and Marcus Rohrbach
FiLM: Feature-wise Linear Modulation

\[
FiLM(F_{i,c} | \gamma_{i,c}, \beta_{i,c}) = \gamma_{i,c}F_{i,c} + \beta_{i,c}
\]

\[
\gamma_{i,c} = f_c(x_i) \quad \beta_{i,c} = h_c(x_i)
\]

\(\gamma, \beta\) change how features are used as learned functions of conditioning input \(x_i\)

FiLM: Feature-wise Linear Modulation

FiLM: Feature-wise Linear Modulation

Histogram of FiLM Parameter Values

$\gamma_{i,c}$

$\beta_{i,c}$

t-SNE of FiLM Parameter Values

Equal [Shape/Color/Size/Mat.] ?

What [Shape/Color/Size/Mat.] ?

t-SNE of FiLM Parameter Values

Equal [Shape/Color/Size/Mat.]?

Equal Number of ...?

Fewer/More of ...?

Live Demo

- Example Questions:
  - “What is the shape of the gray matte object to the right of the large ball that is right of the yellow cylinder?”
  - “What number of things are matte objects that are behind the large cube or big purple shiny balls?”
  - “How many…”
  - “What material is…”
  - “Is there…”
  - “Are there more…”
Logical Inconsistency

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>How many gray things are there?</td>
<td>1</td>
</tr>
<tr>
<td>How many cyan things are there?</td>
<td>2</td>
</tr>
<tr>
<td>Are there as many gray things as cyan things?</td>
<td>Yes</td>
</tr>
<tr>
<td>Are there more gray things than cyan things?</td>
<td>No</td>
</tr>
<tr>
<td>Are there fewer gray things than cyan things?</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Zero-Shot Generalization with FiLM

Q: What shape is the... ...purple thing? A: cube
Q: What shape is the... blue thing? A: sphere
Activation Visualizations

Q: What shape is the... red thing right of the blue thing? A: sphere
Q: What shape is the... red thing left of the blue thing? A: cube
Challenges in VQA

• Image representation
• Language representation
• Combining the modalities
• Attention
• Question-specific reasoning
Standard Approach

Figure from de Vries et al. “Modulating early visual processing by language.” arXiv 2017.

Slide credit: Devi Parikh
What are sitting in the basket of a bicycle?
SANs perform multi-step reasoning
1. Question model
2. Image model
3. Multi-level attention model
4. Answer predictor
1. The image model in the SAN

Spatial feature vectors of different regions of the image

$U_i$ 196 vectors (14 x14)
2. The question model in the SAN

Code the question into a vector using an LSTM

What are sitting in the basket on a bicycle?

\[ v_q \]

**Question:** what are ... bicycle
2. The question model in the SAN (alternative)

Code the question into a vector using a CNN

What are sitting in the basket on a bicycle?

$\mathbf{v}_q$
3. SAN: Computing the 1\textsuperscript{st} level attention

\[ v_I = \sum_i p_i v_i \]
\[ u = v_I + v_q \]

Attention 1

Multimodal Pooling (level 1)

Answer: dogs

To the next attention level

Slide credit: Adapted from Xiaodong He
3. SAN: Compute the 2\textsuperscript{nd} level attention

\[ v_I^{(2)} = \sum_i p_i v_i u^{(2)} = v_I^{(2)} + u \]

Attention

Multimodal Pooling (level 2)

Query vector from the 1\textsuperscript{st} level attention

To the answer predictor

Answer: dogs
4. Answer prediction

\[ u^{(2)} \xrightarrow{w} \text{Softmax} \xrightarrow{} \text{Answer} \]

\[ v_i \]

\[ v_q \]

Answer: dogs

Slide credit: Adapted from Xiaodong He
## Results

<table>
<thead>
<tr>
<th>Methods</th>
<th>test-dev</th>
<th>test-std</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Yes/No</td>
</tr>
<tr>
<td><strong>VQA: [1]</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question</td>
<td>48.1</td>
<td>75.7</td>
</tr>
<tr>
<td>Image</td>
<td>28.1</td>
<td>64.0</td>
</tr>
<tr>
<td>Q+I</td>
<td>52.6</td>
<td>75.6</td>
</tr>
<tr>
<td>LSTM Q</td>
<td>48.8</td>
<td>78.2</td>
</tr>
<tr>
<td>LSTM Q+I</td>
<td>53.7</td>
<td>78.9</td>
</tr>
<tr>
<td><strong>SAN(2, CNN)</strong></td>
<td>58.7</td>
<td>79.3</td>
</tr>
</tbody>
</table>

Table 5: VQA results on the official server, in percentage

**Big improvement** on the VQA benchmark (and COCO-QA, DAQUAR) Improvement is mainly in the *Other* category.
## Results

<table>
<thead>
<tr>
<th>Methods</th>
<th>All</th>
<th>Yes/No 36%</th>
<th>Number 10%</th>
<th>Other 54%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAN(1, LSTM)</td>
<td>56.6</td>
<td>78.1</td>
<td>41.6</td>
<td>44.8</td>
</tr>
<tr>
<td>SAN(1, CNN)</td>
<td>56.9</td>
<td>78.8</td>
<td>42.0</td>
<td>45.0</td>
</tr>
<tr>
<td>SAN(2, LSTM)</td>
<td>57.3</td>
<td>78.3</td>
<td><strong>42.2</strong></td>
<td>45.9</td>
</tr>
<tr>
<td>SAN(2, CNN)</td>
<td><strong>57.6</strong></td>
<td>78.6</td>
<td>41.8</td>
<td><strong>46.4</strong></td>
</tr>
</tbody>
</table>

Table 6: VQA results on our partition, in percentage

Using multi-level attentions improve the performance significantly (also mainly in the Other category)
Text Question-Answering
MCB: Attention Visualizations

What is the woman feeding the giraffe?

Carrot
What is her **hairstyle** for the picture?

**Ponytail**
MCB: Attention Visualizations

What color is the chain on the red dress?

Pink

- Correct Attention, Incorrect Fine-grained Recognition

Slide credit: Akira Fukui and Marcus Rohrbach
MCB: Attention Visualizations

Is the man going to fall down?

No
What is the surface of the court made of?

Clay
MCB: Attention Visualizations

What **sport** is being played?

**Tennis**
MCB: Attention Visualizations

What does the shop sell?

Clocks

• Incorrect Attention
What credit card company is on the banner in the background? 

Budweiser

- Correct Attention, Incorrect Concept Association
Challenges in VQA

• Image representation
• Language representation
• Combining the modalities
• Attention
• Question-specific reasoning
Neural Module Network (NMN) [Andreas 2016]
What color is the necktie?

yellow
What rivers are in South Carolina?

<table>
<thead>
<tr>
<th>name</th>
<th>type</th>
<th>coastal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Columbia</td>
<td>city</td>
<td>no</td>
</tr>
<tr>
<td>Cooper</td>
<td>river</td>
<td>yes</td>
</tr>
<tr>
<td>Charleston</td>
<td>city</td>
<td>yes</td>
</tr>
</tbody>
</table>

Cooper
Grounded question answering

Is there a red shape above a circle? 

yes
Neural nets learn lexical groundings

Is there a red shape above a circle? yes

[Iyyer et al. 2014, Bordes et al. 2014, Yang et al. 2015, Malinowski et al., 2015]
Semantic parsers learn composition

Is there a red shape above a circle?  

[Yes]  


Slide credit: Jacob Andreas
Neural module networks learn both!

*Is there a red shape above a circle?*  

**yes**
Neural module networks

Is there a red shape above a circle?

Slide credit: Jacob Andreas
Neural module networks

Is there a red shape above a circle?

red

exists

above

true

Slide credit: Jacob Andreas
Neural module networks

Is there a red shape above a circle?

red
.exists
.above

→
true
exists

→
red
exists
.above

→
yes

Slide credit: Jacob Andreas
Is there a red shape above a circle?
Representing meaning

Is there a red shape above a circle?

Slide credit: Jacob Andreas
Sets encode meaning

Is there a red shape above a circle?
Sets encode meaning

Is there a red shape above a circle?
Set transformations encode meaning

Is there a red shape above a circle?
Set transformations encode meaning

Is there a red shape above a circle?
Is there a red shape above a circle?
Composing vector functions

exists \quad \text{and} \quad \text{true} \quad \mapsto

red \quad \text{above} \quad \text{circle}
Composing vector functions

Slide credit: Jacob Andreas
Composing vector functions

- red
- exists
- above

→

→ true

→ exists and red above circle

Slide credit: Jacob Andreas
Compositions of vector functions are neural nets
Compositions of vector functions are neural nets
What modules do we need?

Is there a red shape above a circle?

What color is the triangle?

How many goats are there?

What cities are south of San Diego?
Module inventory

Is there a red shape above a circle?

What color is the triangle?

Who is running in the grass?

What cities are south of San Diego?
Learning

Is there a red shape above a circle?

What color is the shape right of a circle?

Slide credit: Jacob Andreas
Is there a red shape above a circle?

What color is the shape right of a circle?
Is there a red shape above a circle?

What color is the shape right of a circle?

Parameter tying

Slide credit: Jacob Andreas
Parameter tying

Is there a red shape above a circle?

What color is the shape right of a circle?
Extreme parameter tying

Slide credit: Jacob Andreas
Is there a red shape above a circle?
Where do layouts come from?

Is there a red shape above a circle?
Choosing among layouts

Is there a red shape above a circle?

Slide credit: Jacob Andreas
Learning to choose layouts

Is there a red shape above a circle?

chooser

Slide credit: Jacob Andreas
Learning with unknown layouts uses RL

Is there a red shape above a circle?

[Williams 1992]
Experiments

<table>
<thead>
<tr>
<th>name</th>
<th>type</th>
<th>coastal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Columbia</td>
<td>city</td>
<td>no</td>
</tr>
<tr>
<td>Cooper</td>
<td>river</td>
<td>yes</td>
</tr>
<tr>
<td>Charleston</td>
<td>city</td>
<td>yes</td>
</tr>
</tbody>
</table>

Slide credit: Jacob Andreas
Experiments: VQA dataset

What color is the necktie? yellow

What is in the sheep’s ear? tag

[Antol et al. 2015]

Slide credit: Jacob Andreas
Experiments: VQA dataset

- Zhou (2015): 55.9
- Noh (2015): 57.4
- NMN: 59.4

Slide credit: Jacob Andreas
Experiments: SHAPES dataset

Slide credit: Jacob Andreas
Experiments: VQA dataset

What color is she wearing?

white
What color is she wearing?
Experiments: VQA Dataset

What is in the sheep’s ear?
What is in the sheep's ear?
What is in the sheep’s ear?
Neural module networks

Linguistic structure dynamically generates model structure

Combines advantages of:

- Representation learning (like a neural net)
- Compositionality (like a semantic parser)
Datasets

Models

Current Status

Ongoing Efforts
Visual Question Answering Challenge 2020

Ayush Shrivastava (Georgia Tech)
Yash Goyal (Georgia Tech → SAIL Montreal)
Dhruv Batra (Georgia Tech / FAIR)
Devi Parikh (Georgia Tech / FAIR)
Aishwarya Agrawal (DeepMind)
Challenge Results
Challenge Results

![Bar chart showing challenge results. The bars represent different groups with varying overall accuracy. The 2018 winner is highlighted in green.]
Challenge Results

+1.10% absolute
Statistical Significance

• Performed Wilcoxon signed-rank test
• @ 95% confidence
Statistical Significance
Easy and Difficult Questions

61% question answered by all top-10 teams

12.8% question not answered by any of top-10 teams
Difficult Questions in 2019 (not in 2020)
Are models sensitive to subtle changes in images?

Who is wearing glasses?

Similar images

man

Different answers

woman
Are models sensitive to subtle changes in images?

• Are predictions accurate for complementary images?

• Accuracy computed for each complementary pair:
  – 1 point: Predict correct answers for both images
  – 0 point, otherwise
Are predictions **accurate** for complementary images?
Are predictions **accurate** for complementary images?
Are models driven by priors?

Non-1-Prior:
- Questions whose answers are not top-1 most common for the given question n-gram in training.
- Consists of 73% of all test-challenge questions.

Agrawal et al., CVPR 2018
Are models driven by priors?
Are models driven by priors?
Are models compositional?

Only consider those questions which are compositionally novel:

- QA pair is not seen in training
- Constituting concepts seen in training

Agrawal et al., Arxiv 2018
Are models compositional?

Training

Q: What color is the plate?
A: Green

Q: What color are stop lights?
A: Red

Testing

Q: What color is the stop light?
A: Green

Q: What is the color of the plate?
A: Red
Are models compositional?

11-13% drop
Are models compositional?
Progress in VQA
Progress in VQA

Human Performance

Accuracy on V2

80
75
70
65
60
55

ICCV 2015

80.78%
Progress in VQA

Human Performance

80.78%

Accuracy on v2

Challenge 2016 Winner

ICCV 2015
Progress in VQA

Human Performance

80.78%

Challenge 2016 Winner

+7.0% absolute
Progress in VQA

Human Performance

80.78%

Accuracy on V2

Challenge 2017 Winner

Challenge 2016 Winner

ICCV 2015
Progress in VQA

Human Performance

80.78%

Accuracy on v2

+6.7% absolute

Challenge 2017 Winner

Challenge 2016 Winner

ICCV 2015

01/01/16 07/01/16 01/01/17 07/01/17 01/01/18 07/01/18 01/01/19 07/01/19 01/01/20
Progress in VQA

Human Performance

80.78%
Progress in VQA

Human Performance

80.78%

+3.4% absolute

Challenge 2017 Winner

Challenge 2018 Winner

ICCV 2015

01/01/16 07/01/16 01/01/17 07/01/17 01/01/18 07/01/18 01/01/19 07/01/19 01/01/20

Accuracy on V2
Progress in VQA

[Graph showing the progress of VQA with markers for ICCV 2015, Challenge 2016 Winner, Challenge 2017 Winner, Challenge 2018 Winner, and Challenge 2019 Winner.]
Progress in VQA

Human Performance

80.78%

+2.9% absolute

Challenge 2016 Winner

Challenge 2017 Winner

Challenge 2018 Winner

Challenge 2019 Winner

ICCV 2015
Progress in VQA

Human Performance

80.78%

Challenge 2020 Winner
Challenge 2019 Winner
Challenge 2018 Winner
Challenge 2017 Winner
Challenge 2016 Winner
ICCV 2015
Progress in VQA
Future Directions

• Very large language models as common-sense priors

• Recent GPT-3 model from OpenAI
  – 175B parameters, ~0.5T words
After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church’s annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.

The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church’s history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.