Visual Question and Answering
Lecture 7

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

Slide credit: Devi Parikh
Visual Question Answering (VQA)

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

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

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

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
Datasets
Models
Current Status
Ongoing Efforts
Datasets
Models
Current Status
Ongoing Efforts
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>set</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>
DAQUAR [Malinowski 2014]

• 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.

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 t-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
FM-IQA [Gao 2015]

- Free-style Multilingual Image Question Answering
- 158,392 COCO images
- 316,193 free-style QAs
- Chinese and English translations

The bus is red.
What is the color of the bus?
香蕉。
What is there in yellow?
羊。
What is there on the grass, except the person?
在椅子上。
Where is the kitty?
西兰花。
Please look carefully and tell me what is the name of the vegetables in the plate?
FM-IQA [Gao 2015]

- FM-IQA
- Gao 2015

**Questions and Answers:**

1. **What is the boy in green cap doing?**
   - He is playing skateboard.

2. **Is there any person in the image?**
   - Yes.

3. **Is the computer on the right hand or left hand side of the gentleman?**
   - On the right hand side.

4. **What is the color of the frisbee?**
   - Yellow.

5. **Why does the bus park there?**
   - Preparing for repair.

6. **What is the texture of the sofa in the room?**
   - Cloth.

7. **Is the man trying to buy vegetables?**
   - Yes.

8. **How many layers are there for the cake?**
   - Six.

9. **What are the people doing?**
   - Walking with umbrellas.
FM-IQA [Gao 2015]

• Evaluation: human
Visual7W [Zhu 2016]

Where does this scene take place?
A) In the sea. ✅
B) In the desert.
C) In the forest.
D) On a lawn.

What is the dog doing?
A) Surfing. ✅
B) Sleeping.
C) Running.
D) Eating.

Why is there foam?
A) Because of a wave. ✅
B) Because of a boat.
C) Because of a fire.
D) Because of a leak.

What is the dog standing on?
A) On a surfboard. ✅
B) On a table.
C) On a garage.
D) On a ball.

Which paw is lifted?
Visual7W [Zhu 2016]

• What, where, when, who, why, how, which
• Telling QA and Pointing QA
• Statistics
  – 47,300 COCO images
  – 327,939 QA pairs
  – 1,311,756 multiple choices
  – 561,459 groundings
Visual7W [Zhu 2016]

Q: What endangered animal is featured on the truck?
A: A bald eagle.
A: A sparrow.
A: A humming bird.
A: A raven.

Q: Where will the driver go if turning right?
A: Onto 24 ¾ Rd.
A: Onto 25 ¾ Rd.
A: Onto 23 ¾ Rd.
A: Onto Main Street.

Q: When was the picture taken?
A: During a wedding.
A: During a bar mitzvah.
A: During a funeral.
A: During a Sunday church service.

Q: Who is under the umbrella?
A: Two women.
A: A child.
A: An old man.
A: A husband and a wife.

Q: Why was the hand of the woman over the left shoulder of the man?
A: They were together and engaging in affection.
A: The woman was trying to get the man’s attention.
A: The woman was trying to scare the man.
A: The woman was holding on to the man for balance.

Q: How many magnets are on the bottom of the fridge?
A: 5.
A: 2.
A: 3.
A: 4.

Q: Which pillow is farther from the window?
Q: Which step leads to the tub?
Q: Which is the small computer in the corner?
Q: Which item is used to cut items?
Q: Which doughnut has multicolored sprinkles?
Q: Which man is wearing the red tie?
Visual7W [Zhu 2016]

- Evaluation: Accuracy
Visual Madlibs [Yu 2015]

**Type 1: image's scene**
- Describe the type of scene/place shown in the image.
- The place is a(n) tennis court.

**Type 2: image's emotion**
- Describe the emotional content of this picture.
- When I look at this picture, I feel hungry.

**Type 3: image's interesting**
- Describe the most interesting or unusual aspect of this picture.
- The most interesting aspect of this picture is... (example: the woman holding the kite)

**Type 4: image's past**
- Describe what happened immediately before this picture was taken.
- One or two seconds before this picture was taken, the man was... (example: riding a horse)

**Type 5: image's future**
- Describe what happened immediately after this picture was taken.
- One or two seconds after this picture was taken, the man was... (example: walking along the street)

**Type 6: object's attribute**
- Describe the appearance of the indicated object.
- The car is white.
Visual Madlibs [Yu 2015]

Type 7: object's affordance
Describe the function of the indicated obj
- People could relax on the couches.

Type 8: object's position
Describe the position of the indicated obj
- The bicycle is in front of the bus.

Type 9: person's attribute
Describe the appearance of the indicated person
- Person A is a balding male.

Type 10: person's activity
Describe the activity of the indicated person
- Person D is standing around.

Type 11: person's location
Describe the location of the indicated person
- Person B is next to an elephant.

Type 12: pair's relationship
Describe the relationship between the individuals
- The person is putting food in the bowl.
Visual Madlibs [Yu 2015]

Task 1: Fill-in-the-blank description

Fill in the blank to describe the activity of the indicated person.
- The person is ____.

Task 2: Multiple-choice question-answering

The most interesting aspect of this picture is ____.
- the biker's position
- the body of the bicycle
- the blue motorcycle
- the child
Visual Madlibs [Yu 2015]

- 10,738 images
- 360,001 descriptions
VQA Dataset
>0.25 million images
254,721 images (COCO)
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!

- 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
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
Plausible Answers

Q. What is he playing?

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

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

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>Age Group</th>
<th>Percentage</th>
<th>Question 1</th>
<th>Question 2</th>
<th>Question 3</th>
<th>Question 4</th>
<th>Question 5</th>
<th>Question 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-4</td>
<td>15.3%</td>
<td>Is that a bird in the sky?</td>
<td>What color is the shoe?</td>
<td>How many zebras are there?</td>
<td>Is there food on the table?</td>
<td>Is this man wearing shoes?</td>
<td></td>
</tr>
<tr>
<td>5-8</td>
<td>39.7%</td>
<td>How many pizzas are shown?</td>
<td>What are the sheep eating?</td>
<td>What color is his hair?</td>
<td>What sport is being played?</td>
<td>Name one ingredient in the skillet.</td>
<td></td>
</tr>
<tr>
<td>9-12</td>
<td>28.4%</td>
<td>Where was this picture taken?</td>
<td>What ceremony does the cake commemorate?</td>
<td>Are these boats too tall to fit under the bridge?</td>
<td>What is the name of the white shape under the batter?</td>
<td>Is this at the stadium?</td>
<td></td>
</tr>
<tr>
<td>13-17</td>
<td>11.2%</td>
<td>Is he likely to get mugged if he walked down a dark alleyway like this?</td>
<td>Is this a vegetarian meal?</td>
<td>What type of beverage is in the glass?</td>
<td>Can you name the performer in the purple costume?</td>
<td>Besides these humans, what other animals eat here?</td>
<td></td>
</tr>
<tr>
<td>18+</td>
<td>5.5%</td>
<td>What type of architecture is this?</td>
<td>Is this a Flemish bricklaying pattern?</td>
<td>How many calories are in this pizza?</td>
<td>What government document is needed to partake in this activity?</td>
<td>What is the make and model of this vehicle?</td>
<td></td>
</tr>
<tr>
<td>Question</td>
<td>Average Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>-------------</td>
<td></td>
<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>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>why</td>
<td>11.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what type</td>
<td>11.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what kind</td>
<td>10.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is this</td>
<td>10.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what does</td>
<td>10.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what time</td>
<td>9.81</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>who</td>
<td>9.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>where</td>
<td>9.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>which</td>
<td>9.32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>does</td>
<td>9.29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>do</td>
<td>9.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what is</td>
<td>9.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what are</td>
<td>9.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>are</td>
<td>8.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is the</td>
<td>8.52</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is there</td>
<td>8.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what sport</td>
<td>8.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>how many</td>
<td>7.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what animal</td>
<td>6.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what color</td>
<td>6.6</td>
<td></td>
<td></td>
<td></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
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?”

2×2×512 LSTM

1024

Point-wise multiplication

Fully-Connected

Softmax over 1000 most frequent answers

“2”
Ablation #2: Vision-alone

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

2x2x512 LSTM

“How many horses are in this image?”

2 x 2 x 512 LSTM

“2”

Point-wise multiplication

Fully-Connected

Softmax over 1000 most frequent answers
<table>
<thead>
<tr>
<th></th>
<th>Open-Answer</th>
<th></th>
<th>Multiple-Choice</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>Yes/No</td>
<td>Number</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>50.39</td>
<td>78.41</td>
<td>34.68</td>
<td>30.03</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>
</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>

“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>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>prior (&quot;yes&quot;)</td>
<td>29.66</td>
<td>70.81</td>
<td>0.39</td>
<td>0.15</td>
<td>29.66</td>
<td>70.81</td>
</tr>
<tr>
<td>per Q-type prior</td>
<td>37.54</td>
<td>71.03</td>
<td>35.77</td>
<td>0.938</td>
<td>39.45</td>
<td>71.02</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>
</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>0.42</td>
<td>0.377</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><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>
</tr>
<tr>
<td>Caption</td>
<td>26.70</td>
<td>65.50</td>
<td>0.02</td>
<td>0.36</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>
Demo

Ask any question about this image

www.visualqa.org
Malinowski 2015
Challenges in VQA

• Image representation
• Language representation
• Combining the modalities
• Attention
• Question-specific reasoning
Multimodal Compact Bilinear (MCB) Pooling [Fukui 2016]
How to Combine Image Representation and Question Representation?

Is this going to be a feast?

- All elements can interact
- Multiplicative interaction

Yes
How to Combine Image Representation and Question Representation?

Is this going to be a feast?

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

Is this going to be a feast?

- All elements can interact
- Multiplicative interaction
  - Difficult to learn input embedding

Elementwise Multiplication

Yes

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

Is this going to be a feast?

Is?

feast

going to be

✓ All elements can interact
✓ Multiplicative interaction

Outer Product / Bilinear Pooling [Lin ICCV 2015]

All elements can interact

Multiplicative interaction

[CNN]

spoon
plate
bowl
table
food
corn

... person

Is this going to be a feast?

Yes

[FC]

Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji.

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?

Outer Product / Bilinear Pooling

Is this going to be a feast?

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

CNN

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

2048

4 million

4 million x 1000

FC

Yes

---


Slide credit: Akira Fukui and Marcus Rohrbach
Multimodal Compact Bilinear Pooling

"Is this going to be a feast?"

- All elements can interact
- Multiplicative interaction
- Low #activations & computation
- Low #parameters

[ICLR Workshops 2016] Fine-grained pose prediction, normalization, and recognition
N Zhang, E Shelhamer, Y Gao, T Darrell


Slide credit: Akira Fukui and Marcus Rohrbach
Multimodal Compact Bilinear Pooling

Is this going to be a feast?

CNN

LSTM

Pham & Pagh (2013):
\[ \Psi(x \otimes y) = \Psi(x) \ast \Psi(y) \]

Random Projection: Countsketch \( \Psi \)

✓ All elements can interact
✓ Multiplicative interaction
☐ Low #activations & computation
✓ Low #parameters

16k x 1000


Slide credit: Akira Fukui and Marcus Rohrbach
Multimodal Compact Bilinear Pooling

Is this going to be a feast?

Pham & Pagh (2013):
\[ \Psi(x \times y) = \Psi(x) \ast \Psi(y) \]

- All elements can interact
- Multiplicative interaction
- Low #activations & computation
- Low #parameters


Slide credit: Akira Fukui and Marcus Rohrbach
Multimodal Compact Bilinear Pooling

Is this going to be a feast?

Yes

All elements can interact
Multiplicative interaction
Low #activations & computation
Low #parameters

Pham & Pagh (2013):
\[ \Psi(x \otimes y) = \Psi(x) \ast \Psi(y) \]


Slide credit: Akira Fukui and Marcus Rohrbach
Experimental setup (without Attention)

- **Solver**
  - Cross-entropy-loss, Adam, learning rate 0.0007

- **Feature Extraction**
  - ResNet 152, image: 448x448

- **Answers**
  - 3000 most frequent on train
  - Sampling with probability of answers

- **Trained on train / validated on val / tested on test-dev**

Slide credit: Devi Parikh
Ablation Comparison to other multimodal methods

- MCB achieves highest accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>Trained on train, test-dev Acc. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eltwise Sum</td>
<td>56.5</td>
</tr>
<tr>
<td>Concat</td>
<td>57.5</td>
</tr>
<tr>
<td>Concat + FC</td>
<td>58.4</td>
</tr>
<tr>
<td>Concat + FC + FC</td>
<td>57.1</td>
</tr>
<tr>
<td>Eltwise Product</td>
<td>58.6</td>
</tr>
<tr>
<td>Eltwise Product + FC</td>
<td>56.4</td>
</tr>
<tr>
<td>Eltwise Product + FC + FC</td>
<td>57.8</td>
</tr>
<tr>
<td>MCB (2048x2048 -&gt; 16k)</td>
<td>59.8</td>
</tr>
<tr>
<td>Full Bilinear (128x128 -&gt; 16k)</td>
<td>58.5</td>
</tr>
<tr>
<td>MCB (128x128 -&gt; 4k)</td>
<td>58.7</td>
</tr>
</tbody>
</table>

Slide credit: Akira Fukui and Marcus Rohrbach
Ablation Comparison to other multimodal methods

• MCB comparable to Full Bilinear
Techniques to improve performance

- Data Augmentation
  - VQA data from Visual Genome Dataset
    - Additional 1M Question and answer pairs
    - Removed articles, Single word answer
- Ensembles
  - Average the output of Softmax over models

VQA Open-Ended accuracy for genome and ensemble

- MCB + Attention (train) 62.5
- MCB + Attention (train + val) 64.2
- MCB + Attention (train + val + genome) 65.1
- MCB + Attention + Ensemble (train + val + genome) 66.7
- MCB + Attention + Ensemble (train + val + genome) Multiple Choice 70.2

Slide credit: Akira Fukui and Marcus Rohrbach
MCB on other Datasets and Tasks

• Visual 7w (Multiple Choice) • Visual Grounding

Our architecture for Visual 7w: MCB with Attention and Answer Encoding.

Accuracy on Visual7W

- Zhu et al.: 54.3
- Concat + Attention: 52.8
- MCB + Attention: 62.2

Accuracy on Flickr30k Entities

- Plummer et al.: 43.8
- Wang et al.: 43.9
- Rohrbach et al.: 47.7
- Concat: 46.5
- Eltwise Prod: 47.4
- Eltwise Prod + Conv: 47.9
- MCB: 48.7

A dog distracts his owner from working at her computer.

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

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 ... ?

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

Activation Visualizations

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

Q: What shape is the ... blue thing? A: sphere
Q: *What shape is the red thing right of the blue thing? A: sphere*
Activation Visualizations

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.
[Yang 2016] Stacked Attention Network (SAN)

Original Image  First Attention Layer  Second Attention Layer

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)

Slide credit: Adapted from Xiaodong He
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?

Slide credit: Adapted from Xiaodong He
2. The question model in the SAN (alternative)

Code the question into a vector using a CNN

Slide credit: Xiaodong He
3. SAN: Computing the 1\textsuperscript{st} level attention

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

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)

To the answer predictor
4. Answer prediction

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

\[ V_i \]

\[ V_q \]

\[ \text{Softmax} \rightarrow \text{Answer: dogs} \]
# 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>VQA: [1]</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>SAN(2, CNN)</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>42.2</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

- John moved to the garden.
- John got the apple there.
- John moved to the kitchen.
- Sandra got the milk there.
- John dropped the apple.
- John moved to the office.

Input Module

Answer
- Kitchen

Episodic Memory

Attention Mechanism
Memory Update

[Xiong 2016]
MCB: Attention Visualizations

What is the woman feeding the giraffe?

Carrot
What is her hairstyle for the picture?

Ponytail
What color is the chain on the red dress? 

Pink

• Correct Attention, Incorrect Fine-grained Recognition
MCB: Attention Visualizations

Is the man going to **fall down**?

No
MCB: Attention Visualizations

What is the surface of the court made of?

Clay
What sport is being played?

Tennis
MCB: Attention Visualizations

What does the shop sell?

Clocks

- Incorrect Attention
MCB: Attention Visualizations

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
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?


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?

Red
Exists
Above
Arrow
True
Is there a red shape above a circle?

- **red**
- **exists**
- **above**

→ red
→ exists
→ above

→ true
Neural module networks

Is there a red shape above a circle?

red

exists

above

→

true

→

→

yes

Slide credit: Jacob Andreas
Representing meaning

Is there a red shape above a circle?
Representing meaning

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

Is there a red shape above a circle?

Slide credit: Jacob Andreas
Sets encode meaning

Is there a red shape above a circle?

Slide credit: Jacob Andreas
Set transformations encode meaning

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

Is there a red shape above a circle?
Sentence meanings are computations

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

Slide credit: Jacob Andreas
Composing vector functions

Slide credit: Jacob Andreas
Composing vector functions

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?

Slide credit: Jacob Andreas
Parameter tying

Is there a red shape above a circle?

What color is the shape right of a circle?
Parameter tying

Is there a red shape above a circle?

What color is the shape right of a circle?

Slide credit: Jacob Andreas
Extreme parameter tying
Where do layouts come from?

Is there a red shape above a circle?

```
be

shape

red  above

circle
```
Where do layouts come from?

Is there a red shape above a circle?

Slide credit: Jacob Andreas
Choosing among layouts

Is there a red shape above a circle?

 chooser

Slide credit: Jacob Andreas
Learning to choose layouts

Is there a red shape above a circle?

chooser
Learning with unknown layouts uses RL

Is there a red shape above a circle?

[Williams 1992]

Slide credit: Jacob Andreas
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>
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

*Zhou

*Yang

NMN
What color is she wearing?

- color
- wear
- white
Experiments: VQA Dataset

What color is she wearing?

slide credit: Jacob Andreas
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
VQA Challenge @ CVPR16

VQA Real Image Challenge (Open-Ended)
Organized by vqateam - Current server time: March 22, 2016, 5 a.m. UTC

Current
Real Challenge test2015 (oe)
Oct. 21, 2015, midnight UTC

Next
Real test2015 (oe)
Oct. 21, 2015, midnight UTC

Overview
Evaluation
Terms and Conditions

Visual Question Answering (VQA)

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), where the goal is to...

Slide credit: Devi Parikh
VQA Challenge @ CVPR16

Current state of the art (MCB)
Open-ended: 66%
Multiple-choice: 70%
Real Open-Ended Challenge

Overall Accuracy

UC Berkeley & Sony
Naver Labs
DLAI T
snubi-naverlabs
POSTECH
Brandeis
VTComputerVision
MIL-UT
klab
SHB_1026
MMCX
VT_CSViasen
LV-NUS
ACVT_Adelaidc
(0MN)
CNNAtt
san
UC Berkeley (NMN)
global_vision
Mujtaba hasan
RIT
UPV UB
Bolei
att
vqateam-lstm cnn
UPC

arXiv v6

ICCV15

Slide credit: Aishwarya Agrawal
Real Open-Ended Challenge

Overall Accuracy

+12.76% absolute

UC Berkeley & Sony
Naver Labs
DLAIT
snubi-naerlabs
POSTECH
Brandeis
VTComputerVison
MIL-UT
klab
SHB_1026
MMCX
VT_CV_Jiasen
LV-NUS
ACVT_Adeelaide
UC Berkeley (DNMN)
CNNAtt
san
UC Berkeley (NMN)
global_Vision
Mujtaba hasan
RIT
UPV_UB
Bolei
att
vqateam-lstm_cnn
UPC
Statistical Significance

- Bootstrap samples 5000 times
- @ 99% confidence