Visual Dialog

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, Dhruv Batra

Presented by: Alan Luo
Introduction **Natural Language Processing + Computer Vision**

- Aiding visually impaired users in understanding their surroundings or social media content
- Interacting with an AI assistant
Related Work  Image/Video Captioning

Image Captioning

1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

Video Captioning

Input video:

Our output: A cat is playing with a toy.

Humans: A Ferret and cat fighting with each other. / A cat and a ferret are playing. / A kitten is playing with a ferret. / A kitten and a ferret are playfully wrestling.

a man is throwing a frisbee in a park

a man riding a wave on top of a surfboard
Related Work  Visual-Semantic Alignments

Visual-Semantic Alignments

Datasets

<table>
<thead>
<tr>
<th>Regions</th>
<th>Attributes</th>
<th>Relationships</th>
</tr>
</thead>
<tbody>
<tr>
<td>the white bowls on the table</td>
<td>bowl is white</td>
<td>top</td>
</tr>
<tr>
<td>the woman teaching the little girl to cook</td>
<td>pan is silver</td>
<td>lady Wearing shirt</td>
</tr>
<tr>
<td>the liquid in the bowl</td>
<td>kitchen utensils is kitchen</td>
<td>woman helping girl</td>
</tr>
<tr>
<td>the utensil on the food</td>
<td>kitchen utensils is behind</td>
<td>sauce on bread</td>
</tr>
<tr>
<td>the pink shirt on the woman</td>
<td>utensils is behind</td>
<td>tin foil under bread</td>
</tr>
<tr>
<td>the pink shirt on the little girl</td>
<td>pizza is unripe</td>
<td>paper hanging on wall</td>
</tr>
<tr>
<td>little girl with a pink top on next to</td>
<td>hand is white</td>
<td>lady helping girl</td>
</tr>
<tr>
<td>Question Answers</td>
<td>shirt is pink</td>
<td>child standing at table</td>
</tr>
<tr>
<td></td>
<td>girl is young</td>
<td></td>
</tr>
<tr>
<td></td>
<td>spoon is large</td>
<td>woman IN kitchen</td>
</tr>
<tr>
<td></td>
<td>spoon is silver</td>
<td></td>
</tr>
</tbody>
</table>

What color is the girl's shirt?  Pink
Where were the photo taken?  Kitchen
What are the people doing?  Cooking
How many adults in the kitchen?  One
What color is the sauce being put onto the food?  Red
Related Work Visual Q&A

- How many pickles are on the plate?
  - 1
- What is the shape of the plate?
  - circle
  - round

- What does the sign say?
  - stop
  - yield
- What shape is this sign?
  - octagon
  - diamond
  - round
Contributions

1. Propose a new AI task: Visual Dialog
2. Develop a novel two-person chat data-collection protocol and introduce a new dataset
3. Introduce a family of neural encoder-decoder models for Visual Dialog
Technical Details With Late Fusion Encoder

Image 1

Do you think the woman is with him?

Question $Q_t$

The man is riding his bicycle on the sidewalk. Is the man wearing a helmet? No he does not have a helmet on. ... Are there any people nearby? Yes there's a woman walking behind him.

t rounds of history (concatenated)

Decoder

No I don't think they are together

Answer $A_t$
Dataset **VisDial**

**Qualitative**

Caption: A statue depicting a bear breaking into a car.
Person A (1): how big is statue
Person B (1): about size of real full grown bear
Person A (2): so is car full size then as well
Person B (2): yes replica of car
Person A (3): is statue all 1 color
Person B (3): no brown and black
Person A (4): what color is car
Person B (4): dark red
Person A (5): where is this, do you think
Person B (5): in wooded area someplace
Person A (6): do you see any people in image
Person B (6): yes 1 man
Person A (7): how old is man
Person B (7): 35-40
Person A (8): what is man doing
Person B (8): sitting in car behind replica
Person A (9): do you see any signs
Person B (9): yes, on car door warning sign
Person A (10): what else can you tell me about this image
Person B (10): there are many trees in background

**Quantitative**

- Questions vs. Answers
- VQA vs. Visual Dialog

![Graphs showing percentage coverage and unique answers](image_url)
Qualitative Results

Caption: A group of bikers parked in a parking lot under large lights
moped
on street
cloudy sky looks clear

Quantitative Results

<table>
<thead>
<tr>
<th>Model</th>
<th>MRR</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer prior</td>
<td>0.311</td>
<td>19.85</td>
<td>39.14</td>
<td>44.28</td>
<td>31.56</td>
</tr>
<tr>
<td>NN-Q</td>
<td>0.392</td>
<td>30.54</td>
<td>46.99</td>
<td>49.98</td>
<td>30.88</td>
</tr>
<tr>
<td>NN-QI</td>
<td>0.385</td>
<td>29.71</td>
<td>46.57</td>
<td>49.86</td>
<td>30.90</td>
</tr>
<tr>
<td>LF-Q-G</td>
<td>0.403</td>
<td>29.74</td>
<td>50.10</td>
<td>56.32</td>
<td>24.06</td>
</tr>
<tr>
<td>LF-Q1-H-G</td>
<td>0.425</td>
<td>32.49</td>
<td>51.56</td>
<td>57.80</td>
<td>23.11</td>
</tr>
<tr>
<td>LF-Q1-I-G</td>
<td>0.437</td>
<td>34.06</td>
<td>52.50</td>
<td>58.89</td>
<td>22.31</td>
</tr>
<tr>
<td>LF-Q1-H-G</td>
<td>0.430</td>
<td>33.27</td>
<td>51.96</td>
<td>58.09</td>
<td>23.04</td>
</tr>
<tr>
<td>HRE-Q1-H-G</td>
<td>0.430</td>
<td>32.84</td>
<td>52.36</td>
<td>58.64</td>
<td>22.59</td>
</tr>
<tr>
<td>HRE-Q1-I-G</td>
<td>0.442</td>
<td>34.37</td>
<td>53.40</td>
<td>59.74</td>
<td>21.75</td>
</tr>
<tr>
<td>HRE-Q1-H-I</td>
<td>0.442</td>
<td>34.47</td>
<td>53.43</td>
<td>59.73</td>
<td>21.83</td>
</tr>
<tr>
<td>MN-Q1-H-G</td>
<td>0.434</td>
<td>35.12</td>
<td>55.14</td>
<td>59.61</td>
<td>22.14</td>
</tr>
<tr>
<td>MN-Q1-H-I</td>
<td>0.443</td>
<td>34.62</td>
<td>53.74</td>
<td>60.18</td>
<td>21.69</td>
</tr>
</tbody>
</table>

What is this?
Where is this?
How is the weather?