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

Presented by Wei-Chieh Wu
Visual Dialog

• Requires an AI agent to hold a meaningful dialog with humans about visual content.

• Input:
  • Image
  • Dialog history
  • Question

• Output:
  • Answer to the question
VQA vs Visual Dialog

**VQA**
Q: How many people on wheelchairs?
A: Two

Q: How many wheelchairs?
A: One

**Captioning**
Two people are in a wheelchair and one is holding a racket.

**Visual Dialog**
Q: How many people are on wheelchairs?
A: Two

Q: What are their genders?
A: One male and one female

Q: Which one is holding a racket?
A: The woman

**Visual Dialog**
Q: What is the gender of the one in the white shirt?
A: She is a woman

Q: What is she doing?
A: Playing a Wii game

Q: Is that a man to her right?
A: No, it's a woman
VisDial Dataset

- Contains \(~123k\) images and 10 question-answer pairs for each image
- Images are from COCO dataset
- Question-answer pairs are collected on AMT with human dialog
VisDial Dataset
VisDial Dataset

(a) VisDial Questions
(b) VQA Questions
(c) VisDial Answers
Evaluation

• Given N = 100 candidate answers, return a sorting of them

• Candidate answers:
  • The human response
  • Answers to 50 most similar questions
  • 30 most popular answers from the dataset
  • 19 random answers

• Retrieval metrics:
  MRR, recall@k, average rank of the human response
Models

• Following the encoder-decoder framework
• 2 kinds of decoder
  - Generative Decoder
  - Discriminative Decoder
• 3 kinds of encoder
  - Late Fusion Encoder
  - Hierarchical Recurrent Encoder
  - Memory Network Encoder
Decoders

• Generative Decoder
  • LSTM decoder
  • Maximize the log-likelihood of the ground truth answer
  • Use the model’s log-likelihood scores for ranking

• Discriminative Decoder
  • Compute similarity between the input encoding and LSTM encoding for candidate answers
  • Maximize softmax score of the ground truth answer
  • Use the similarities for ranking
Late Fusion (LF) 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)
Hierarchical Recurrent Encoder (HRE)
Memory Network (MN) Encoder

Image I

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.
- How old is the man? He looks around 40 years old.
- What color is his bike? It has black wheels and handlebars, I can't see the body of the bike that well.
- Is anyone else riding a bike? No he's the only one.
- Are there any people nearby? Yes there's a woman walking behind him.

$t$ rounds of history

$\{(\text{Caption}), (Q_t, A_t), \ldots, (Q_t, A_t)\}$
Experiments

• Dataset: VisDial v0.9

• Baseline
  • NN-Q:
    Find k nearest neighbor questions for a test question, and score answers by their mean similarity with these k answers
  • NN-QI:
    Find K nearest neighbor questions for a test question, then find a subset of size k based on image feature similarity. Score answers by their mean similarity with these k answers

• VQA models
  • SAN and HieCoAtt
  • Feed VQA outputs to their discriminative decoder, and train end-to-end on VisDial
<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><strong>Baseline</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Answer prior</td>
<td>0.3735</td>
<td>23.55</td>
<td>48.52</td>
<td>53.23</td>
<td>26.50</td>
</tr>
<tr>
<td>NN-Q</td>
<td>0.4570</td>
<td>35.93</td>
<td>54.07</td>
<td>60.26</td>
<td>18.93</td>
</tr>
<tr>
<td>NN-QI</td>
<td>0.4274</td>
<td>33.13</td>
<td>50.83</td>
<td>58.69</td>
<td>19.62</td>
</tr>
<tr>
<td><strong>Generative</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LF-Q-G</td>
<td>0.5048</td>
<td>39.78</td>
<td>60.58</td>
<td>66.33</td>
<td>17.89</td>
</tr>
<tr>
<td>LF-QH-G</td>
<td>0.5055</td>
<td>39.73</td>
<td>60.86</td>
<td>66.68</td>
<td>17.78</td>
</tr>
<tr>
<td>LF-QI-G</td>
<td>0.5204</td>
<td>42.04</td>
<td>61.65</td>
<td>67.66</td>
<td>16.84</td>
</tr>
<tr>
<td>LF-QIH-G</td>
<td>0.5199</td>
<td>41.83</td>
<td>61.78</td>
<td>67.59</td>
<td>17.07</td>
</tr>
<tr>
<td>HRE-QH-G</td>
<td>0.5102</td>
<td>40.15</td>
<td>61.59</td>
<td>67.36</td>
<td>17.47</td>
</tr>
<tr>
<td>HRE-QIH-G</td>
<td>0.5237</td>
<td><strong>42.29</strong></td>
<td>62.18</td>
<td>67.92</td>
<td>17.07</td>
</tr>
<tr>
<td>HREA-QIH-G</td>
<td>0.5242</td>
<td>42.28</td>
<td>62.33</td>
<td>68.17</td>
<td><strong>16.79</strong></td>
</tr>
<tr>
<td>MN-QH-G</td>
<td>0.5115</td>
<td>40.42</td>
<td>61.57</td>
<td>67.44</td>
<td>17.74</td>
</tr>
<tr>
<td>MN-QIH-G</td>
<td><strong>0.5259</strong></td>
<td><strong>42.29</strong></td>
<td><strong>62.85</strong></td>
<td><strong>68.88</strong></td>
<td>17.06</td>
</tr>
<tr>
<td><strong>Discriminative</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LF-Q-D</td>
<td>0.5508</td>
<td>41.24</td>
<td>70.45</td>
<td>79.83</td>
<td>7.08</td>
</tr>
<tr>
<td>LF-QH-D</td>
<td>0.5578</td>
<td>41.75</td>
<td>71.45</td>
<td>80.94</td>
<td>6.74</td>
</tr>
<tr>
<td>LF-QI-D</td>
<td>0.5759</td>
<td>43.33</td>
<td>74.27</td>
<td>83.68</td>
<td>5.87</td>
</tr>
<tr>
<td>LF-QIH-D</td>
<td>0.5807</td>
<td>43.82</td>
<td>74.68</td>
<td>84.07</td>
<td>5.78</td>
</tr>
<tr>
<td>HRE-QH-D</td>
<td>0.5695</td>
<td>42.70</td>
<td>73.25</td>
<td>82.97</td>
<td>6.11</td>
</tr>
<tr>
<td>HRE-QIH-D</td>
<td>0.5846</td>
<td>44.67</td>
<td>74.50</td>
<td>84.22</td>
<td>5.72</td>
</tr>
<tr>
<td>HREA-QIH-D</td>
<td>0.5868</td>
<td>44.82</td>
<td>74.81</td>
<td>84.36</td>
<td>5.66</td>
</tr>
<tr>
<td>MN-QH-D</td>
<td>0.5849</td>
<td>44.03</td>
<td>75.26</td>
<td>84.49</td>
<td>5.68</td>
</tr>
<tr>
<td>MN-QIH-D</td>
<td><strong>0.5965</strong></td>
<td><strong>45.55</strong></td>
<td><strong>76.22</strong></td>
<td><strong>85.37</strong></td>
<td><strong>5.46</strong></td>
</tr>
<tr>
<td><strong>VQA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAN1-QI-D</td>
<td>0.5764</td>
<td>43.44</td>
<td>74.26</td>
<td>83.72</td>
<td>5.88</td>
</tr>
<tr>
<td>HieCoAtt-QI-D</td>
<td>0.5788</td>
<td>43.51</td>
<td>74.49</td>
<td>83.96</td>
<td>5.84</td>
</tr>
</tbody>
</table>