VQA:
Visual Question Answering

Group No: 02

E/15/043  Bhagya T.P.Y.
E/15/092  Ekanayake I.U.
E/15/187  Kulanjith G.D.
E/15/246  Opanayake R.L.
PRESENTATION OUTLINE

Overview of the research
Introduction
Related work
Dataset
Dataset Analysis
VAQ baseline and methods

KEY DISCUSSION POINTS
WHAT?

THEY HAVE DONE?

Open-ended Q&A
- Complex reasoning & detailed understanding
- Images=0.25M  Q=0.75  A=10.0M
- Small Questions and Closed set of Answers ("yes" or "no" or small 1 to 3 words answers)
INTRODUCTION

Multi-discipline Artificial Intelligence

- Computer Vision (CV)
- Natural Language Processing (NLP)
- Knowledge Representation & Reasoning (KR)

What is AI-complete?

Combination of human understanding and computer technology

multi-modal knowledge + quantitative evaluation metric
INTRODUCTION

Type of Answers
- Open-ended answering
- Multiple-choice

Evaluation?
Number of questions it answers correctly

Datasets
MS COCO - 204,721 images
abstract scene dataset - 50,000 scenes (3Qs)
RELATED WORK

Other VQA works

- M. Malinowski and M. Fritz - (Small data set/ Small range of Questions)
- D. Geman and the team (A Visual Turing Test for Computer Vision Systems)
- K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S. C. Zhu (Video VQA)
  By providing a text and a video answer

Not open-ended, Not free-form Qs & As
RELATED WORK

- Sentence completion with multiple-choice answers.
- Grounding of questions (understand text + image).
- Commonsense, knowledge, and complex reasoning.

VQA

- Text-based Q&A
- Describing Visual Content
- Vision+Language Tasks

- Image tagging
- Image captioning
- Video captioning (Generate sentences to describe visual content.)
- Detailed specific info

Evaluate image captioning
- Coreference resolution
- Generating referring expressions
DATASETS

- MS COCO - 204,721 images
- Abstract scene dataset - 50,000 scenes
The MS COCO dataset already contains five single-sentence captions for all images.

Abstract scene dataset
- 20 "paperdoll" human models spanning genders, races, and ages with 8 different expressions
- 100 objects and 31 animals in various poses
Collecting Questions

- Simple questions - require low-level computer vision knowledge. 
  ex- “What color is the cat?”
- Questions that require commonsense knowledge about the scene. 
  ex- “What sound does the pictured animal make?”
- Three questions for each image/scene.
- Dataset contains over ~0.76M questions.
Collecting Answers

• Open-ended questions result in a diverse set of possible answers.
• 10 answers for each question from unique workers.
Testing

Accuracy metric:
\[
\min\left( \frac{\# \text{ humans that provided that answer}}{3}, 1 \right)
\]

- 100% accuracy if at least 3 workers provided that exact answer.
VAQ Dataset Analysis

• Provide an analysis of the questions and answers in the VQA train dataset
  • To gain an understanding of the types of questions asked and answers provided, the following things can be done:
    • Visualize the distribution of question types and answers
    • Explore how often the questions may be answered without the image using just common sense information
    • Analyze whether the information contained in an image caption is sufficient to answer the questions
Types of Questions

Real Images

Abstract Scenes
Answers

- Typical Answers
- Lengths
- 'Yes/No' and 'Number' Answers
• Subject Confidence

• Inter-human Agreement
• COMMON SENSE of KNOWLEDGE

• Is the Image Necessary?

e.g. - What is the colour of a fire hydrant?

<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 BASELINES AND METHODS

Baselines

- random: randomly choose an answer from the top 1K answers of the VQA train/val dataset
- prior ("yes"): always select the most popular answer ("yes") for both the open-ended and multiple-choice tasks.
- per Q-type prior:
  For the open-ended task: pick the most popular answer per question type.
  For the multiple-choice task: pick the answer that is most similar to the picked answer in the open-ended task (cosine similarity in Word2Vec feature space)
- k nearest neighbor
Methods

2-channel vision (image) + language (question) model
Image Channel: This channel provides an embedding for the image.

1. The activations from the last hidden layer of VGGNet are used as 4096-dim image embedding.
2. norm I: These are l2 normalized activations from the last hidden layer of VGGNet.

Question Channel: This channel provides an embedding for the question.

1. Bag-of-Words Question (BoW Q)
2. LSTM Q
3. deeper LSTM Q
**Results**

<table>
<thead>
<tr>
<th></th>
<th>Open-Ended</th>
<th></th>
<th>Multiple-Choice</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></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>
</tr>
<tr>
<td>per Q-type prior</td>
<td>37.54</td>
<td>71.03</td>
<td>35.77</td>
<td>09.38</td>
</tr>
<tr>
<td>nearest neighbor</td>
<td>42.70</td>
<td>71.89</td>
<td>24.36</td>
<td>21.94</td>
</tr>
<tr>
<td>BoW Q</td>
<td>48.09</td>
<td>75.66</td>
<td>36.70</td>
<td>27.14</td>
</tr>
<tr>
<td>I</td>
<td>28.13</td>
<td>64.01</td>
<td>00.42</td>
<td>03.77</td>
</tr>
<tr>
<td>BoW 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>48.76</td>
<td>78.20</td>
<td>35.68</td>
<td>26.59</td>
</tr>
<tr>
<td>LSTM Q + I</td>
<td>53.74</td>
<td>78.94</td>
<td>35.24</td>
<td>36.42</td>
</tr>
<tr>
<td>deeper LSTM Q</td>
<td>50.39</td>
<td>78.41</td>
<td>34.68</td>
<td>30.03</td>
</tr>
<tr>
<td>deeper LSTM Q + norm I</td>
<td>57.75</td>
<td><strong>80.50</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>BoW Q + C</td>
<td>54.70</td>
<td>75.82</td>
<td>40.12</td>
<td>42.56</td>
</tr>
</tbody>
</table>

**TABLE 2:** Accuracy of our methods for the open-ended and multiple-choice tasks on the VQA test-dev for real images. Q = Question, I = Image, C = Caption. (Caption and BoW Q + C results are on val. See text for details.)

vision-alone model that completely ignores the question performs rather poorly best model (deeper LSTM Q + norm I)
Conclusion

- Large data set is used providing more generalization to the VQA Task
- Data obtained from real persons
- Contribution to the idea of "Ai complete"
- For some applications Task specific question may improve performance
THANK YOU!