Visual Question Answering

Liang-Wei Chen, Shuai Tang
Outline

- Problem statement
- Common VQA benchmark datasets
- Methods
- Dataset bias problem
- Future Development
Problem Statement
What is VQA?

Given an image, can our machine answer the corresponding questions in natural language?

VQA Demo
How to see and how to read

**Question:** What can the red object on the ground be used for?

**Answer:** Firefighting

How to see and how to read


**Question:** What can the red object on the ground be used for?

**Answer:** Firefighting
How to see and how to read

Knowledge Representation and reasoning

Computer Vision

Natural Language Understanding

Question: What can the red object on the ground be used for?
Answer: Firefighting
Support Fact: Fire hydrant can be used for fighting fires.

Multiple choices V.S. Open-ended settings

How many cats are there?

(a) One
(b) Two
(c) Three
(d) Four

Common VQA Benchmark Datasets
VQA dataset

VQA-real

Q: Where are the magazines in this picture?
A: On stool, stool, on stool, on bar stool, on table, stool, on stool, on chair, on bar stool, stool

VQA-abstract

Q: Where are the flowers?
A: near tree, tree, around tree, tree, by tree, around tree, around tree, grass, beneath tree, base of tree

VQA dataset

- Offer open-ended answers and multiple choice answers
- 250k images (MS COCO + 50k abstract images)
- 750k questions, 10M answers
- Each question is answered by 10 human annotators

Microsoft COCO-QA

- Automatically generate QA pairs with MS COCO captions
- 123,287 images (72,783 for training and 38,948 for testing) and each image has one QA pair.
- 4 types of question templates: What object, How many, What color, Where

**Visual7W**

- 47,300 COCO images, 327,939 QA pairs, and 1,311,756 human-generated multiple-choices
- 7W stands for what, where, when, who, why, how and which

---

Telling

Pointing

---

Evaluation metrics

Recall metrics for image captioning, BLEU, etc.
- Emphasize on similarity between sentences.

Major VQA datasets use Accuracy:
- Exact string matching.
- Most answers are 1 to 3 words.

VQA dataset (with 10 human annotations):

\[
\text{accuracy} = \min \left( \frac{\text{# humans provided that answer}}{3}, 1 \right)
\]
## A brief summary over VQA datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of QA pairs</th>
<th>Annotation</th>
<th>Question Diversity</th>
<th>Answer Diversity (top-1000 answers coverage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>COCO-QA</td>
<td>117,684</td>
<td>Open-ended</td>
<td>Low (generated from captions)</td>
<td>100%</td>
</tr>
<tr>
<td>VQA</td>
<td>614,163</td>
<td>Open-ended +Multiple choices</td>
<td>Encourage to have diverse annotations</td>
<td>82.7%</td>
</tr>
<tr>
<td>Visual 7W</td>
<td>327,939</td>
<td>Multiple choices</td>
<td>At least 3 W for each image</td>
<td>63.5%</td>
</tr>
</tbody>
</table>

Methods

Language-Image Embedding
Bag-of-words + Image feature (iBOWIMG)

are these people family?

Bag-of-words + Image feature (iBOWIMG)

Bag-of-words + Image feature (iBOWIMG)

LSTM + Image feature (LSTM Q + I)

“How many horses are in this image?”

Performances on the VQA-real dataset (test-dev)

<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>Overall</td>
<td>Yes/No</td>
<td>Number</td>
<td>Other</td>
<td>Overall</td>
<td>Yes/No</td>
</tr>
<tr>
<td>iBOWIMG</td>
<td>55.72</td>
<td>76.55</td>
<td>35.03</td>
<td>42.62</td>
<td>61.68</td>
<td>76.68</td>
</tr>
<tr>
<td>LSTM Q+I</td>
<td>57.75</td>
<td>80.50</td>
<td>36.77</td>
<td>43.08</td>
<td>62.70</td>
<td>80.52</td>
</tr>
</tbody>
</table>

Multimodal Compact Bilinear Pooling (MCB)

- Use outer product for embedding.
- Results in high dimensional features.
- Then approximate it with low dimensional features.

Multimodal Compact Bilinear Pooling (MCB)

Bilinear Pooling

Multimodal Compact Bilinear Pooling (MCB)

Compact Bilinear Pooling via Count Sketch

Multimodal Compact Bilinear Pooling (MCB)

MCB+soft attention for VQA open-ended questions:

Multimodal Compact Bilinear Pooling (MCB)

MCB+soft attention for VQA open-ended questions:

Multimodal Compact Bilinear Pooling (MCB)

MCB+soft attention for VQA open-ended questions:

Multimodal Compact Bilinear Pooling (MCB)

Answer Encoding
(for multiple choice questions):

Q: “What do you see?” (Ground Truth: a₃)
   a₁: “A courtyard with flowers”
   a₂: “A restaurant kitchen”
   a₃: “A family with a stroller, tables for dining”
   a₄: “People waiting on a train”

Multimodal Compact Bilinear Pooling (MCB)

Comparison of multimodal pooling methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Element-wise Sum</td>
<td>56.50</td>
</tr>
<tr>
<td>Concatenation</td>
<td>57.49</td>
</tr>
<tr>
<td>Concatenation + FC</td>
<td>58.40</td>
</tr>
<tr>
<td>Concatenation + FC + FC</td>
<td>57.10</td>
</tr>
<tr>
<td>Element-wise Product</td>
<td>58.57</td>
</tr>
<tr>
<td>Element-wise Product + FC</td>
<td>56.44</td>
</tr>
<tr>
<td>Element-wise Product + FC + FC</td>
<td>57.88</td>
</tr>
<tr>
<td>MCB (2048 $\times$ 2048 $\rightarrow$ 16K)</td>
<td><strong>59.83</strong></td>
</tr>
</tbody>
</table>
## Multimodal Compact Bilinear Pooling (MCB)

### Additional Experiment results

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Bilinear (128 x 128 → 16K)</td>
<td>58.46</td>
</tr>
<tr>
<td>MCB (128 x 128 → 4K)</td>
<td>58.69</td>
</tr>
<tr>
<td>Element-wise Product with VGG-19</td>
<td>55.97</td>
</tr>
<tr>
<td>MCB (d = 16K) with VGG-19</td>
<td><strong>57.05</strong></td>
</tr>
<tr>
<td>Concatenation + FC with Attention</td>
<td>58.36</td>
</tr>
<tr>
<td>MCB (d = 16K) with Attention</td>
<td><strong>62.50</strong></td>
</tr>
</tbody>
</table>

Methods

1. Region-based Image Attention
Focus on image regions to answer questions?

Is it raining?  

What color is the walk light?

Extract top ranked 99 regions from edge box + (1 whole image)
Features are from ImageNet by VGGnets


Co-attention of image and question

\[
X_r = \begin{pmatrix}
\text{region1} \\
\text{region2} \\
\vdots \\
\text{regionN}
\end{pmatrix}^T
\]

\[
G_r = AX_r + b_r
\]

\[
\vec{s}_{l,r} = \sigma(G_r^T \vec{g}_l)
\]

(softmax)

\[
\vec{x}_l = \text{text features}
\]

(word embedding)

\[
\vec{g}_l = B\vec{x}_l + b_l
\]

Linearly combine the region features

What color on the stop light is lit up?

L: red (-0.1)
I: red (-0.8)
R: green (1.1)
Ans: green

What color is the light?

L: red (1.0)
I: red (0.3)
R: red (1.7)
Ans: red

What color is the street sign?

L: gray (-0.2)
I: gray (-0.4)
R: yellow (0.4)
Ans: yellow
Performances on the VQA-real dataset (test-dev)

<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>Overall</td>
<td>Yes/No</td>
<td>Number</td>
<td>Other</td>
<td>Overall</td>
<td>Yes/No</td>
<td>Number</td>
<td>other</td>
</tr>
<tr>
<td>iBOWIMG</td>
<td>55.72</td>
<td>76.55</td>
<td>35.03</td>
<td>42.62</td>
<td>61.68</td>
<td>76.68</td>
<td>37.05</td>
<td>54.44</td>
</tr>
<tr>
<td>LSTM Q+I</td>
<td>57.75</td>
<td>80.50</td>
<td>36.77</td>
<td>43.08</td>
<td>62.70</td>
<td>80.52</td>
<td>38.22</td>
<td>53.01</td>
</tr>
<tr>
<td>Region attention</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>62.44</td>
<td>77.62</td>
<td>34.28</td>
<td>55.84</td>
</tr>
</tbody>
</table>
### Accuracies by type of question

- **On the VQA-real dataset (validation set)**

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Region</th>
<th>Image</th>
<th>Text</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>58.94</td>
<td>57.83</td>
<td>53.98</td>
<td>100.0%</td>
</tr>
<tr>
<td>is/are/was</td>
<td>75.42</td>
<td>74.63</td>
<td>75.00</td>
<td>33.3%</td>
</tr>
<tr>
<td>Identify: what kind/type/animal</td>
<td>52.89</td>
<td>52.10</td>
<td>45.11</td>
<td>23.8%</td>
</tr>
<tr>
<td>How many</td>
<td>33.38</td>
<td><strong>36.84</strong></td>
<td>34.05</td>
<td>10.3%</td>
</tr>
<tr>
<td>What color</td>
<td><strong>53.96</strong></td>
<td>43.52</td>
<td>32.59</td>
<td>9.8%</td>
</tr>
<tr>
<td>Interpret: can/could/does/has</td>
<td>75.73</td>
<td>74.43</td>
<td><strong>75.73</strong></td>
<td>4.6%</td>
</tr>
<tr>
<td>None of the above</td>
<td>45.40</td>
<td>44.04</td>
<td><strong>48.23</strong></td>
<td>4.1%</td>
</tr>
<tr>
<td>Where</td>
<td>42.11</td>
<td><strong>42.43</strong></td>
<td>37.61</td>
<td>2.5%</td>
</tr>
<tr>
<td>Why/How</td>
<td>26.31</td>
<td>28.18</td>
<td><strong>29.24</strong></td>
<td>2.2%</td>
</tr>
</tbody>
</table>

- **Region**: region weighted features
- **Image**: whole image features
- **Text**: no image feature

Methods

2. Hierarchical Question Attention
Hierarchical Question-Image Co-Attention (HieCoAtt)

Question: What color on the stop light is lit up?

Answer: green

Image

Question hierarchy

"What color on the ... up?"

Word embedding

Question hierarchy

Question hierarchy

Question Level

Word Level

The colored words are those with higher weights.

The colored words are those with higher weights.

Ablation study on the VQA-real dataset (validation set)

- The attention mechanisms closest to the ‘top’ of the hierarchy matter most

<table>
<thead>
<tr>
<th>Method</th>
<th>validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y/N</td>
</tr>
<tr>
<td>LSTM Q+I</td>
<td>79.8</td>
</tr>
<tr>
<td>Image Atten</td>
<td>79.8</td>
</tr>
<tr>
<td>Question Atten</td>
<td>79.4</td>
</tr>
<tr>
<td>W/O Q-Atten</td>
<td>79.6</td>
</tr>
<tr>
<td>W/O P-Atten</td>
<td>79.5</td>
</tr>
<tr>
<td>W/O W-Atten</td>
<td>79.6</td>
</tr>
<tr>
<td>Full Model</td>
<td>79.6</td>
</tr>
</tbody>
</table>

## Performances on the VQA-real dataset (test-dev)

<table>
<thead>
<tr>
<th>Method</th>
<th>Open-Ended Overall</th>
<th>Open-Ended Yes/No</th>
<th>Open-Ended Number</th>
<th>Open-Ended Other</th>
<th>Multiple-Choice Overall</th>
<th>Multiple-Choice Yes/No</th>
<th>Multiple-Choice Number</th>
<th>Multiple-Choice other</th>
</tr>
</thead>
<tbody>
<tr>
<td>iBOWIMG</td>
<td>55.72</td>
<td>76.55</td>
<td>35.03</td>
<td>42.62</td>
<td>61.68</td>
<td>76.68</td>
<td>37.05</td>
<td>54.44</td>
</tr>
<tr>
<td>LSTM Q+I</td>
<td>57.75</td>
<td>80.50</td>
<td>36.77</td>
<td>43.08</td>
<td>62.70</td>
<td>80.52</td>
<td>38.22</td>
<td>53.01</td>
</tr>
<tr>
<td>MCB</td>
<td>60.8</td>
<td>81.2</td>
<td>35.1</td>
<td>49.3</td>
<td>65.40</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Region attention</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>62.44</td>
<td>77.62</td>
<td>34.28</td>
<td>55.84</td>
</tr>
<tr>
<td>HieCoAtt</td>
<td>61.80</td>
<td>79.78</td>
<td>38.78</td>
<td>51.78</td>
<td>65.80</td>
<td>79.70</td>
<td>40.00</td>
<td>59.80</td>
</tr>
</tbody>
</table>
Dataset Bias Problem
Baselines that Exploit Dataset Biases

By predicting correctness of an Image-Question-Answer triplet:

- reaches state-of-the-art performance on Visual7W Telling.
- performs competitively on VQA Real multiple choice.

Models:

\[ \text{MLP: } y = \sigma(W_2 \max(0, W_1 x_{iqa}) + b) \]
Baselines that Exploit Dataset Biases

Baselines:
- MLP(A,Q,I): use all features
- MLP(A,I): Answers + Images
- MLP(A,Q): Answers + Questions
- MLP(A): Answers

Accuracies on Visual7W Telling

<table>
<thead>
<tr>
<th>Method</th>
<th>What</th>
<th>Where</th>
<th>When</th>
<th>Who</th>
<th>Why</th>
<th>How</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM (Q, I) [15]</td>
<td>48.9</td>
<td>54.4</td>
<td>71.3</td>
<td>58.1</td>
<td>51.3</td>
<td>50.3</td>
<td>52.1</td>
</tr>
<tr>
<td>MCB + Att [21]</td>
<td>60.3</td>
<td>70.4</td>
<td>79.5</td>
<td>69.2</td>
<td>58.2</td>
<td>51.1</td>
<td>62.2</td>
</tr>
<tr>
<td>MLP (A)</td>
<td>47.3</td>
<td>58.2</td>
<td>74.3</td>
<td>63.6</td>
<td>57.1</td>
<td>49.6</td>
<td>52.9</td>
</tr>
<tr>
<td>MLP (A, Q)</td>
<td>54.9</td>
<td>60.0</td>
<td>76.8</td>
<td>66.0</td>
<td>64.5</td>
<td>54.9</td>
<td>58.5</td>
</tr>
<tr>
<td>MLP (A, I)</td>
<td>60.8</td>
<td>74.9</td>
<td>81.9</td>
<td>70.3</td>
<td>64.4</td>
<td>51.2</td>
<td>63.8</td>
</tr>
<tr>
<td>MLP (A, Q, I)</td>
<td>64.5</td>
<td>75.9</td>
<td>82.1</td>
<td>72.9</td>
<td>68.0</td>
<td>56.4</td>
<td>67.1</td>
</tr>
</tbody>
</table>

**Baselines that Exploit Dataset Biases**

Accuracies on VQA Multiple Choice (test)

<table>
<thead>
<tr>
<th>Method</th>
<th>Yes/No</th>
<th>Number</th>
<th>Other</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-Layer LSTM [5]</td>
<td>80.6</td>
<td>37.7</td>
<td>53.6</td>
<td>63.1</td>
</tr>
<tr>
<td>Region selection [23]</td>
<td>77.2</td>
<td>33.5</td>
<td>56.1</td>
<td>62.4</td>
</tr>
<tr>
<td>MCB [21]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>65.4</td>
</tr>
<tr>
<td>MCB + Att + GloVe + Genome [21]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td><strong>69.9</strong></td>
</tr>
<tr>
<td>MLP (A, Q, I)</td>
<td>80.8</td>
<td>17.6</td>
<td><strong>62.0</strong></td>
<td>65.2</td>
</tr>
</tbody>
</table>

* evaluated on test-dev set
Responses to Dataset Biases

1. More Balanced Datasets
CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning

- 100,000 computer rendered images.
- 864,986 generated questions.
- Scene graph representations for all images.
- Functional program representation for all images.

With complex questions like: counting, attribute identification, comparison, multiple attention, and logical operations.

A sample image and questions from CLEVR

VQA 2.0: Making the V in VQA matter

Who is wearing glasses?
- man
- woman

Where is the child sitting?
- fridge
- arms

Is the umbrella upside down?
- yes
- no

How many children are in the bed?
- 2
- 1

VQA 2.0: Making the V mater

Results:

VQA models trained/tested on unbalanced/balanced datasets

<table>
<thead>
<tr>
<th>Approach</th>
<th>UU(%)</th>
<th>UB(%)</th>
<th>BB(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language Only</td>
<td>48.21</td>
<td>40.03</td>
<td>39.98</td>
</tr>
<tr>
<td>LSTM(Q+I)</td>
<td>54.40</td>
<td>46.56</td>
<td>48.18</td>
</tr>
<tr>
<td>HieCoAtt</td>
<td>57.09</td>
<td>49.51</td>
<td>51.02</td>
</tr>
<tr>
<td>MCB</td>
<td>60.36</td>
<td>53.67</td>
<td>55.35</td>
</tr>
</tbody>
</table>
Responses to Dataset Biases

2. Compositional Models
Neural Module Network (NMN)

Exploit compositional nature of questions:

<table>
<thead>
<tr>
<th>Question</th>
<th>Code</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>how many different lights in various different shapes and sizes?</td>
<td>describe<a href="find%5Blight%5D">count</a></td>
<td>four (four)</td>
</tr>
<tr>
<td>what is the color of the horse?</td>
<td>describe<a href="find%5Bhorse%5D">color</a></td>
<td>brown (brown)</td>
</tr>
<tr>
<td>what color is the vase?</td>
<td>describe<a href="find%5Bvase%5D">color</a></td>
<td>green (green)</td>
</tr>
<tr>
<td>is the bus full of passengers?</td>
<td>describe[is](combine[and](find[bus], find[full]))</td>
<td>yes (yes)</td>
</tr>
<tr>
<td>is there a red shape above a circle?</td>
<td>measure[is](combine[and](find[red], transform<a href="find%5Bcircle%5D">above</a>))</td>
<td>yes (yes)</td>
</tr>
</tbody>
</table>

What color is his tie?

Is there a red shape above a circle?
Neural Module Network (NMN)

Neural Modules:

Neural Module Network (NMN)

Neural Modules:

Find

Transform

Neural Module Network (NMN)

Neural Modules:

Find

Image $\rightarrow$ Attention

Combine

Attention $\times$ Attention $\rightarrow$ Attention

Transform

Attention $\rightarrow$ Attention

Describe

Image $\times$ Attention $\rightarrow$ Label

Measure

Attention $\rightarrow$ Label

Neural Module Network (NMN)

From questions to networks:

- Parse questions to structured queries:

  “Is there a circle next to a square?” -> is(circle, next-to(square))
Neural Module Network (NMN)

From questions to networks:

- Parse questions to structured queries:
  
  “Is there a circle next to a square?” -> is(circle, next-to(square))

- Generate network Layout:

  Question to layout as a deterministic process

Neural Module Network (NMN)

Final Model:
Why using LSTM feature?

- aggressive simplification of the question, need grammar cues.
- capture semantic regularities, e.g. common sense.

Answering “Where is the dog?” with NMN

## Neural Module Network (NMN)

### Results on VQA

Results on the VQA open ended questions

<table>
<thead>
<tr>
<th>Method</th>
<th>test-dev</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes/No</td>
<td>Number</td>
</tr>
<tr>
<td>LSTM</td>
<td>78.7</td>
<td>36.6</td>
</tr>
<tr>
<td>ATT+LSTM</td>
<td>80.6</td>
<td>36.4</td>
</tr>
<tr>
<td>NMN</td>
<td>70.7</td>
<td>36.8</td>
</tr>
<tr>
<td>NMN+LSTM</td>
<td>81.2</td>
<td>35.2</td>
</tr>
<tr>
<td>NMN+LSTM+FT</td>
<td>81.2</td>
<td>38.0</td>
</tr>
</tbody>
</table>

Generate layout candidates:

The input sentence (a) is represented as a dependency parse (b). Fragments are then associated with modules (c), fragments are assembled into full layouts (d).

Select best candidates:

\[
s(z_i|x) = a^T \sigma (Bh_q(x) + Cf(z_i) + d)
\]

\[
p(z_i|x; \theta_\ell) = e^{s(z_i|x)} / \sum_{j=1}^{n} e^{s(z_j|x)}
\]

What is the constraint in this setting?
Dynamical Neural Module Network (DNMN)

- Evaluate all layouts $p(z|x; \theta_e)$, cheap
- Evaluate all answers given all layout $p_z(y|w; \theta_e)$, expensive

A reinforcement learning problem at heart!

State $\rightarrow (w, x)$  Action $\rightarrow z$  Reward $\rightarrow \log p(y|z, w; \theta_e)$

Policy gradient: $\mathbb{E}[(\nabla \log p(z|x; \theta_e)) \cdot \log p(y|z, w; \theta_e)]$ (REINFORCE rule)

Adding language prior to DNMN model: $\log p_z(y|w, x) = (Ah_q(x) + B[z]_w)_y$

Dynamical Neural Module Network (DNMN)

Results on the VQA open ended questions

<table>
<thead>
<tr>
<th></th>
<th>test-dev</th>
<th></th>
<th></th>
<th>test-std</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes/No</td>
<td>Number</td>
<td>Other</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Zhou (2015)</td>
<td>76.6</td>
<td>35.0</td>
<td>42.6</td>
<td>55.7</td>
<td>55.9</td>
</tr>
<tr>
<td>NMN</td>
<td>81.2</td>
<td>38.0</td>
<td>44.0</td>
<td>58.6</td>
<td>58.7</td>
</tr>
<tr>
<td>D-NMN</td>
<td>81.1</td>
<td>38.6</td>
<td>45.5</td>
<td>59.4</td>
<td><strong>59.4</strong></td>
</tr>
</tbody>
</table>

Zhou (2015) - iBOWING
Future Development
Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

A complete cooperative game of two agents:

**Q-BOT:**
- Build a mental model of the unseen image purely from the natural language dialog,
- Predict an image feature and gain rewards for two agents

**A-BOT:**
- See the target image, generate a caption for Q-BOT
- Build a mental model of what Q-BOT understands, and answer questions.
Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

Visual Dialog

Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

Policy Networks for Q-BOT and A-BOT:

Policy Networks for Q-BOT and A-BOT:

\[ s_t^Q = [c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}] \]

Policy Networks for Q-BOT and A-BOT:

$$s^Q_t = [c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}]$$

$$s^A_t = [I, c, q_1, a_1, \ldots, q_{t-1}, a_{t-1}, q_t]$$

Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning

Results:

(a) Guessing Game Evaluation.

A new language-image task called “Visual Question Answering”
- Can machines see images, reason and answer to questions like human?

Various approaches
- Language-image embedding
- Attention mechanism
- Compositional models

Dataset bias
- Can exploit answer distribution biases
- High question-answer correlation

Further development
- More Balanced dataset
- Emphasis on visual features
Thank you for your attention!
Reading list