Improving VQA and its Explanations by Comparing Competing Explanations

Jialin Wu, Liyan Chen and Raymond J. Mooney
Visual Question Answering (VQA)

• Answering natural language questions about an image

Question: Is this in an Asian country?
Answer: Yes
VQA systems

• Framed as multi-class classification problems

Question: Is this in an Asian country?

Answer: Yes
Lack of Explanatory Capabilities

- Dataset biases introduce shortcuts
  - ‘Is this’ type questions are more likely to be answered as No.
  - ‘Red light’ means No.
  - More US subways are in the dataset.
Lack of Explanatory Capabilities

• Key features are subtle
  • Hard to let the system automatically learn to focus on the text on the train’s marquee, and further notice the type of characters.
VQA with Textual Explanations

• Human annotated explanations

Question: Is this in an Asian country?
Answer: Yes

The information provided on the train’s marquee is comprised of Asian characters.
Competing Explanations

• Plausible explanations that support top-ranked answer candidates.
  • Describe expected behaviors when a certain answer candidate is true.

Candidate 1: No  VQA confidence: 0.88
Sample Retrieved Explanations:
1. The train looks European as well as the railings and surrounding area.
2. The wording on the train is in English.
3. 4…. 8…

Candidate 2: Yes  VQA Confidence: 0.79
Sample Retrieved Explanations:
1. It does not look like a standard American train.
2. The signs are all in Japanese.
3. 4…. 8…
Competing Explanations

• Extracting top-ranked answer candidates
  • We used the top-5 answers from a pretrained VQA systems

• Generating explanations for each answer candidates
  • Retrieval approach: select explanations of the 8 examples in the training set with the most similar features for questions, images and answers.
  • Generation approach: we train an explanation module [1] to generate the explanation for each candidate.

Using Competing Explanations

• Competing explanations are encoded to re-weight the original VQA confidence based on their supportiveness to the answer candidates.
Verification Systems

• score how well a generated or retrieved explanation supports a corresponding answer candidate given the question and visual content.

\[ S(Q, V, a, x) = \sigma(f_2(f(q), f(v), f(a), f(\phi(x)))) \]
Training

• Positive examples
  • The VQA examples with annotated human explanations

• Negative examples
  • Replacing visual features
  • Replacing question features
  • Replacing answer features
  • Replacing explanation features
Experimental Results

• VQA-X Performance pretraining on VQA-X
Experimental Results

- VQA-X
- VQA\textsubscript{v2}
Experimental Results

• VQA-X Performance pretraining on VQA-X
Experimental Results

• VQA-X Performance pretraining on VQA-v2
Experimental Results

• VQA-X Performance pretraining on VQA-v2
Qualitative Results

What type of fruit toy is the cat holding?
Cat(0.0): VQA: 0.12  VQA+E: 0.00
A fluffy animal with ears and a tail is there.
Banana(1.0): VQA: 0.09  VQA+E: 0.08
It is long with a yellow peel.
Qualitative Results

What beverage is in the cup?
Milk(0.0): VQA: 0.20  VQA+E: 0.11
It is a liquid and white.
Beer(1.0): VQA: 0.13  VQA+E: 0.12
It is amber in color.
Competing Explanations for Better Explanation

• Provide references to explanations to similar visual questions.

What is this piece of furniture used for?

Sleep(0.6):
There is a bed and pillows in the room.
Competing Explanations for Better Explanation

• We used the explanation module from [1] and use the retrieved explanations as additional input features.

<table>
<thead>
<tr>
<th></th>
<th>Automatic Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU-4</td>
</tr>
<tr>
<td>Faith. Expl. [1]</td>
<td>25.0</td>
</tr>
<tr>
<td>Faith. Expl. + E (ours)</td>
<td>26.4</td>
</tr>
</tbody>
</table>